AD_____

Award Number: W81XWH-10-1-1006

TITLE: Molecular Profiles for Lung Cancer Pathogenesis and Detection in U.S. Veterans

PRINCIPAL INVESTIGATOR: Steven M. Dubinett, M.D.

CONTRACTING ORGANIZATION: University of California, Los Angeles,
Los Angeles, CA 90095

REPORT DATE: October 2013

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
October 2013	Annual	20 September 2012 – 19 September 2013

4. TITLE AND SUBTITLE

Molecular Profiles for Lung Cancer Pathogensis and Detection in U.S. Veterans

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-10-1-1006

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Steven M. Dubinett, M.D. Pierre Massion, M.D.
Brigitte M. Gompers, M.D.
Ignacio Wistuba, M.D.
Avrum Spria, M.D.
E-Mail: sdubinett@mednet.ucla.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) AND ADDRESS(ES)
University of California, Los Angeles
11000 Kinross Ave., Ste 211
Los Angeles, CA 90095-0001

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
During our first year of research, we demonstrated a localized "field cancerization" phenomenon on gene expression in the airway of patients with lung cancer, and we identified several pathways preferentially activated in the airway adjacent to tumors. In addition, we have identified markers of stem cells in the airway that may represent tumor-initiating cells of the airway and are evaluating profiles of these cells. We have identified Snail as a novel marker of stem cells in the airway that promote epithelial-mesenchymal transition. We have made a major technical advance in developing the methods required to use low quality and quantity laser capture microdissected material to sequence the transcriptome. This allows us to examine the gene expression profiles in premalignant lesions and compare it to the histologically normal airway epithelium and tumor. We have validated this approach and the data will allow us to identify novel pathways for lung carcinogenesis. All of these studies are identifying biomarkers that could be used for early lung cancer detection and pathways that are involved in "field cancerization". Understanding this "field cancerization" and development of premalignant lesions is likely to shed light on novel pathways in lung carcinogenesis that could lead to diagnostic tests, therapies and cancer chemoprevention strategies for lung cancer.

15. SUBJECT TERMS
"field cancerization" ; lung carcinogenesis ; gene expression profiles ; lung cancer stem cells

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
U	U	U	UU	66	19b. TELEPHONE NUMBER *(include area code)*

Table of Contents

Introduction

Lung cancer is the leading cause of death from cancer in the US and the world, accounting for 28% of all cancer deaths in men and women [1]. Lung-cancer associated mortality has remained essentially unchanged over the last 3 decades, in part because the majority of lung cancers present at an advanced stage. Since 1972, when Congress declared war on cancer, lung cancer's 5 year survival rate has remain unchanged at 15% while the 5 year survival rates for breast, prostate and colon cancers have risen to 88%, 99% and 65% respectively [1]. The major cause of lung cancer is smoking, yet only 10-15% of smokers develop lung cancer [2]. Tobacco addiction and exposure to other lung cancer carcinogens are serious problems among military personnel and war veterans. In 2005, the DOD reported that over 32% of military personnel smoke. Although this is a substantial decrease when compared to a survey done in 1980, the smoking rate among the general population in 2005 was only 21% according to a report by the Centers for Disease Control and Prevention. Smoking among military personnel has increased since the late 1990s in association with conflicts in Afghanistan and Iraq and is 50% higher in deployed vs. non-deployed personnel with smoking rates in 20-25 year olds of 37% vs. 20% in the civilian population. Even conservative estimates place the cost of lung cancer to the military at $1 billion a year and this cost will only increase with this wave of new smokers. The optimal treatment for non-small-cell lung cancer (NSCLC) is surgical resection; however, 75% of patients are ineligible because of advanced disease.

The results of National Lung Screening Trial (NLST) that utilized low dose helical computed tomography (LDCT) screening in a population of heavy current or former smokers were recently

published and demonstrated a 20% reduction in lung cancer mortality and a 6.7% decrease in all-cause mortality with LDCT relative to CXR [3-4]. This is the single most significant advance in reducing lung cancer mortality since 1964, when the United States Surgeon General first publicized the causal relationship between lung cancer and cigarette smoking, prompting a decline in active smoking in the US. However, with the reported 96.4% false positives rate there is a pressing need to develop reliable molecular biomarkers to supplement the radiologic imaging to improve the sensitivity of early detection of lung cancer. Based on the notions that: 1) smoking-induced injury alters mRNA and microRNA (miRNA) expression profiles in airway epithelium [5-7], and 2) these changes can be detected and serve as biomarker for early detection of lung cancer [8-9], in the current project we focus on investigation of the molecular events associated with field cancerization to identify biomarkers of early lung carcinogenesis.

In Specific Aim 1 of this program, high-throughput microarray mRNA expression analyses have been performed on cytological specimens (brushings) obtained at intraoperative bronchoscopy from the main carina and main ipsilateral bronchus, as well as on specimens obtained at lobectomy procedures from the main lobe bronchus (adjacent to SCCs), sub-segmental bronchus (adjacent to adenocarcinomas) and from the resected NSCLC tumors. Towards this aim, we are comparing and contrasting global gene expression patterns across all the specimens from the entire field and corresponding NSCLC tumors. We are deriving lung adenocarcinoma and SCC field cancerization signatures signifying the differential mRNA expression patterns between the carina and the subsegmental bronchus and main lobe bronchus, respectively. In addition, similar expression profiles between the carina and resected NSCLC tumors are being integrated with available gene expression data of bronchial brushings from the main carina isolated at various time points post-surgery from 40 NSCLC patients; Department of Defense (DoD) VITAL patients. Promising markers derived from this study are being validated at the mRNA and protein level in histological tissue specimens. Moreover, we are currently performing RNA-sequencing and microarray profiling of nasal epithelia, airway epithelial cells collected from both bronchoscopy and lobectomy specimens as well as of corresponding tumors (NSCLC patients) or benign lesions (cancer-free individuals).

In Specific Aim 2, we are using laser capture microdissection to obtain specific cell populations (basal cells or type II alveolar cells, depending on the NSCLC histology/location) as well as premalignant lesions and epithelial components of the tumors. These cell populations are being profiled with RNA-seq to determine their gene expression signatures to increase our understanding of premalignancy. We are analyzing the gene expression profiles that are associated with progression from a benign cell population to premalignancy and with progression from a benign cell population to true malignancy.

In Specific Aim 3, we will use expression signatures and biomarkers derived from the results of aims 1 and 2 to develop and test airway-based biomarkers capable of diagnosing lung cancer in current or former smokers using minimally invasive sites.

This report details the progress made during the third year of research.

COMPREHENSIVE ANNUAL PROGRESS REPORT FROM LEAD PI (DR. STEVEN DUBINETT)

Molecular Profiles for Lung Cancer Pathogenesis and Detection in U.S. Veterans

Specific Aim 1: To increase our understanding of the molecular basis of the pathogenesis of lung cancer in the "field cancerization" that develops in current and former smokers.

Summary of Research Findings

A. Collection of airway epithelial samples from both bronchoscopy and lobectomy specimens from smokers with and without lung cancer.

We have recruited 37 subjects undergoing resection of lung tumor or benign lung lesions to collect tissue samples for the studies in Aim 1. From these subjects who were recruited at all 4 participating institutions, we have collected nasal epithelium, proximal and distal bronchial airway epithelium obtained at bronchoscopy (ipsilateral and contralateral to the tumor) as well as the tumor/benign lesion, adjacent normal parenchyma, and sub-segmental bronchial epithelium at time of lobectomy. A summary of subjects is provided in **Table 1.**

Institution	RNA-seq (cases)			Microarray (cases)		
	ADC	SCC	No Cancer	ADC	SCC	No Cancer
MD Anderson	4	2	0	9	5	2
BU	0	1	3	0	2	4
UCLA	1	2	1	3	2	2
Vanderbilt	1	1	1	4	3	1
Subtotals	6	6	5	16	12	9
Total # of cases analyzed (samples)	17 (105)			37 (248)		

Table 1. Molecular mapping of the field of injury in NSCLC and cancer-free patients. *ADC, adenocarcinoma; SCC, squamous cell carcinoma; BU, Boston University; UCLA, University of California Los Angeles.*

B. High throughput gene expression profiling (Aim1A).

Total RNA from all samples was isolated using the miRNeasy kit (Qiagen) and profiled by microarray (using the Affymetrix GeneChipHuman Gene 2.0 ST platform). Microarrays were normalized using the Robust Multiarray Average (RMA) with an Entrez Gene-specific probeset mapping from the Molecular and Behavioral Neuroscience Institute (MBNI) at the University of Michigan (version 16). Outliers were removed based on the Relative Log Expression (RLE) and Normalized Unscaled Standard Error (NUSE) quality metrics and Principal Component Analysis (PCA). At 6 different brush sites, differences between subjects with and without cancer were assessed using a linear model. These sites include distal airway close to or further from the tumor, ipsilateralor contralateral proximal airway, main carina, and nose. Each of these comparisons was used to create a ranked list and gene set using Gene Set Enrichment Analysis (GSEA). **Figure 1** shows the field cancerization effect in which cancer-associated changes in gene

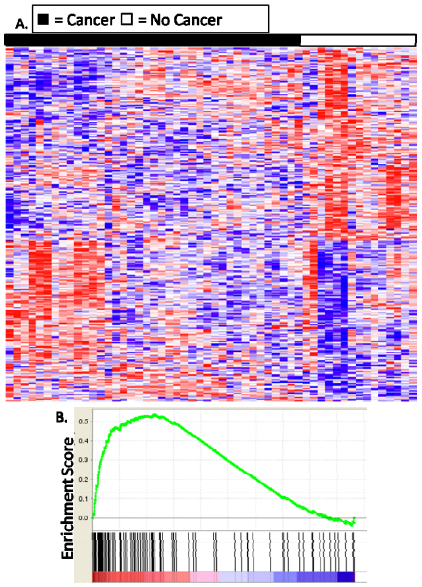

Figure 1: Gene expression patterns in distal and proximal airway. A. We identified differentially expressed genes in cancer vs no cancer throughout the airway, shown above are results for the proximal airway. **B**. GSEA of top 100 genes up in cancer in the proximal airway enriched in a ranked list of cancer genes in the distal airway (fdr < 0.001)

expression in the proximal airway are enriched among genes that change with cancer in the distal airway. Additionally, the cancer-specific airway gene expression changes are enriched among those changing in the lung tumor itself, with the enrichment increasing with proximity to the tumor (**Figure 2).** Gene expression was also profiled using RNA-seq (four-plex, Illumina HiSeq 2000, 100nt paired-end reads), yielding on average 30 million reads per sample. FastQC [http://www.bioinformatics.babraham.ac.uk/projects/fastqc] was used to assess per-base sequence quality and nucleotide composition, and duplication levels. Most samples yielded RNAseq data with mean and median Phred scores above 30 through the length of the read (**Figure 3**). Reads were aligned with TopHat to human genome build 19 (hg19) and assembled with Cufflinks using Ensembl build 64. Novel transcriptome assembly will be completed as follows, using a method developed in our lab. First, TopHat will be run in genome-guided *denovo* mode, and after filtering splice junctions appearing only in one sample, Cuffmerge will be used to combine annotations from all samples. HTseq [http://www-huber.embl.de/users/anders/HTSeq/] will then be used to estimate locus-level count data, and differential expressionof novel transcripts will be assessed using edgeR[15]. We anticipate that analysis of microarray and RNA-seq data will be complete by the end of 2013 and functional analysis completed by Spring 2014.

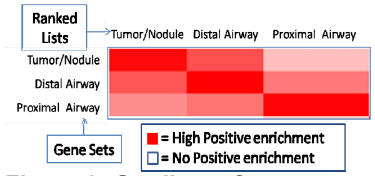

Figure 2: Gradient of gene expression. Analysis of cancer vs no cancer in the proximal airway, distal airway, and tumor (tumor vs benign nodule) was used to generate a ranked list and top 100 genes up in cancer. GSEA was used to compare enrichment across all sites.

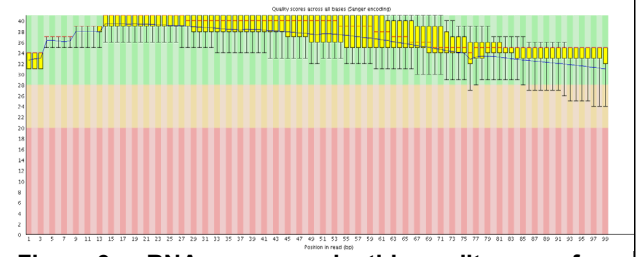

Figure 3: mRNAseq per nucleotide quality score for a representative sample. Quality score at each position is assessed and plotted.

Sequencing of small RNA will be completed by Spring 2014, and sequence read quality will be assessed using FastQC. After removal of adapter sequence, reads will be aligned with Bowtie to hg19 and microRNA expression will be estimated using BEDTools with the most recent build of miRBase. Differential expression of microRNA will be assessed using edgeR, and microRNA-mRNA regulatory relationships will be recovered using tools such as CLR and miRconnX.

This study has, for the first time, allowed us to 1) perform next generation sequencing in addition to microarray profiling analysis of the molecular field of injury in the airway; 2) study samples obtained from four different institutions in the nation using common SOPs and 3) characterize the topological map of the molecular field of injury/cancerization between NSCLC patients and cancer-free individuals. We anticipate that expression profiles in the NSCLC molecular field of injury will harbor transcripts, both novel and established, that may exhibit potential for use as airway biomarkers that can be developed

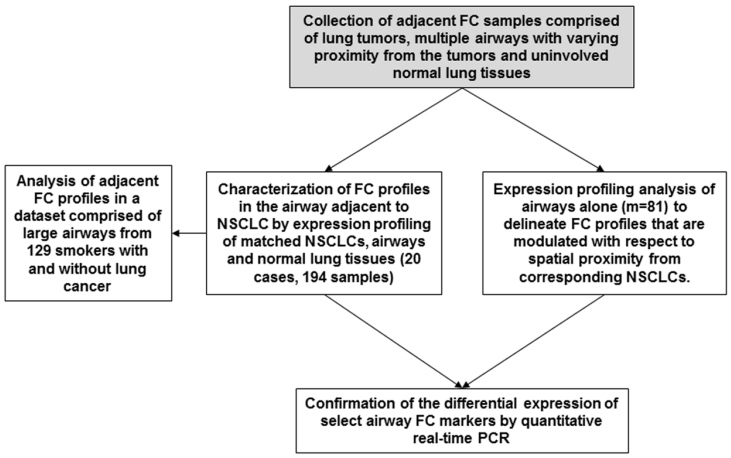

Figure 4. Schematic of the transcriptomic analysis of the airway Field cancerization in NSCLC. Field cancerization samples comprised of matched lung tumors, controlled normal and uninvolved normal lung tissues were obtained from 20 resected early-stage (I-IIIA) NSCLC specimens. Gene expression profiling of all the samples (n=194) was performed to characterize global cancerization profiles in the normal-appearing airway adjacent to NSCLCs. Adjacent field profiles were then analyzed by GSEA in a set of large airways of smokers (n=129) with and without lung cancer from BU (PI: Avrum Spira) to identify profiles that can distinguish lung cancer among smokers. Global expression profiling was also employed to delineate profiles that are modulated in the airway with respect to proximity from corresponding NSCLCs. QRTPCR was used to confirm the differential expression of select airway markers.

and tested for lung cancer detection using minimally invasive sites in Specific Aim 3 of this award.

C. Gene expression analysis of bronchial epithelial samples obtained from lobectomy specimens from NSCLC patients (*Field Cancerization Study*)

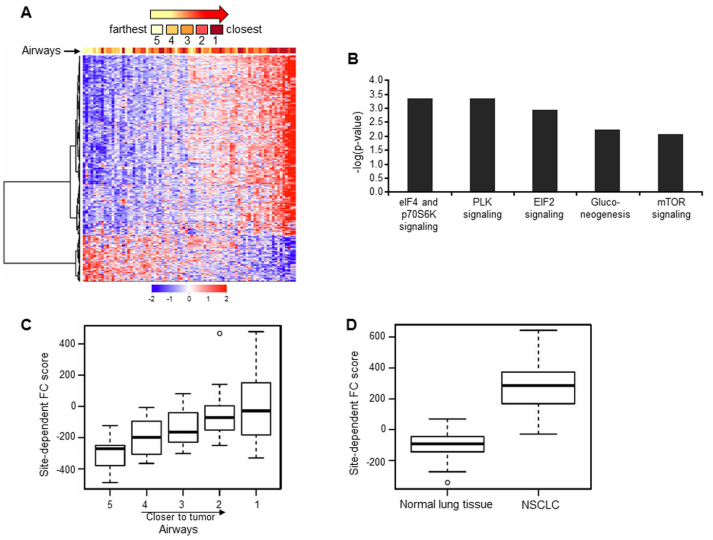

Figure 5. Analysis of airway expression profiles by distance form corresponding NSCLCs. A. Ordinal logistic regression analysis of airways was performed and identified 422 gene features with significant differential expression in the airway with respect to tumor proximity (FDR <5%). **B**. Functional pathways analysis by IPA of the site-dependent differentially expressed genes. Box plots depicting site-dependent FC score in airways (**C**) and between corresponding paired NSCLCs and normal lung tissue (**D**).

We sought to characterize the yet unknown global molecular and adjacent airway field cancerization in early-stage NSCLC. We performed whole-transcriptome expression profiling of resected early-stage (I-IIIA) NSCLC specimens (n=20) with matched tumors, multiple cytologically controlled normal airways with varying distances from tumors and uninvolved normal lung tissues (n=194 samples) using the Affymetrix Human Gene 1.0 ST platform. Samples were obtained from patients who did not receive neoadjuvant therapy undergoing lobectomy or pneumonectomy procedures under an MD Anderson institutional review board (IRB)-approved protocol. Mixed-effects models were used to identify differentially expressed genes among groups. Ordinal regression analysis was performed to characterize site-dependent airway expression profiles. A schematic of the study's design is represented in **Figure 4.** Data from the gene expression analysis have been described and detailed in the previous annual report (Year 2). We identified differentially expressed gene features (n=1661) between NSCLCs and airways compared to normal lung tissues. We then examined the expression of the adjacent airway FC profile in a cohort comprised of 129 large airway samples from smokers with and without lung cancer. This analysis identified a subset (n=299), following gene set enrichment analysis, that significantly (*P* <0.001) and concordantly clustered large airways of healthy smokers from airways of lung cancer patients.

We also identified a cassette of gene features (n=422) that were significantly and progressively differentially expressed in airways by distance from tumors (**Figures 5A and 5B**). Notably, when we examined the 422 gene features in NSCLCs and paired normal lung tissues, we found that 291 of the 335 genes that were increased and 53 of the 87 that were decreased in airways with shorter distance from tumors were also up-regulated and down-regulated, respectively, in NSCLCs compared to normal lung tissues and that a field cancerization score quantifying the site-dependent effect in the airway was congruently modulated between NSCLCs and normal lung (**Figures 5C and 5D**). Our findings suggest that the adjacent airway field of cancerization harbors profiles and pathways that are both site-independent as well as gradient and localized with respect to nearby tumors and that may point to new molecular tools for detection of NSCLC and further inform of the molecular pathology of the malignancy. These data have been submitted as an abstract in the past AACR and have been submitted for publication and currently under revision (**see Reportable Outcomes**).

D. Expression validation and functional studies

We performed qRT-PCR analysis of select field cancerization markers including lysosomal protein transmembrane 4 beta (*LAPTM4B*), which was among the top 5 FC markers with increased expression in airways with respect to tumor proximity. **As detailed in our previous annual report (Year 2)**, QRTPCR confirmed microarray data and demonstrated that 1) *LAPTM4B* was significantly increased in NSCLCs and in airways with shorter distance from tumors; 2) *LAPTM4B* was significantly increased in immortalized lung epithelial cells compared to normal bronchial cells and 3) transient knockdown of *LAPTM4B* expression in immortalized and malignant lung epithelial cell lines significantly reduced cell growth. We sought to further confirm the role of *LAPTM4B* in lung cancer cell growth by generating sub-lines with stable knockdown of the gene in the Calu-6 cell line with high basal level of *LAPTM4B*. Stable knockdown of *LAPTM4B* significantly suppressed *LAPTM4B* (P <0.001) expression (**Figure 6A**) concomitant with significantly reduced cell growth (P <0.05) (**Figure 6B**) and anchorage-dependent (**Figure 6C**) and -independent colony number (all P <0.05) (**Figure 6D**). These data demonstrate that *LAPTM4B* is a positive mediator of malignant lung epithelial cell growth and phenotype. Moreover, these findings are proof of principle that characterization of field cancerization markers can lead to the identification of otherwise unknown mediators of lung cancer molecular pathogenesis. These data have also been submitted as an abstract in the past AACR and have been submitted for publication and currently under revision (**see Reportable Outcomes**).

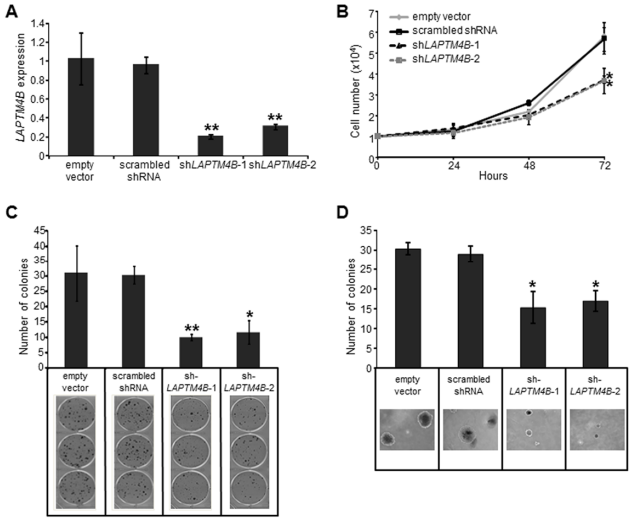

Figure 6. Stable knockdown of *LAPTM4B* reduces lung cancer cell growth and colony formation. A. QRTPCR analysis of Calu-6 cells stably transfected with empty vectors or vectors expressing scrambled shRNA or *LAPTM4B*-specific shRNA. **B**. Trypan blue exclusion count of the stably transfected cells. Anchorage-dependent (**C**) and –independent (**D**) colony formation assay of the stably transfected Calu-6 cells was performed. Representative images of cell colonies are depicted in the lower panels. Error bars indicate standard deviation. **P* <0.05; ***P* <0.001 by the Student's *t* test.

We then sought to examine the expression of *LAPTM4B* in NSCLC formalin-fixed paraffin embedded (FFPE) histological tissue specimens. We opted to assess *LAPTM4B* transcript expression by *in situ* hybridization (ISH) due to limited antibodies available for immunohistochemical analysis of the protein product coding for this relatively understudied gene. We employed the QuantiGene 2.0 kit and QuantiGene View (QGV) RNA ISH tissue assay from Affymetrix (Santa Clara, CA) according to the manufacturer's instructions. The assay comprised singleplex probe for *LAPTM4B* based on the oncogene's transcript reference sequence from the NCBI (NM_018407). We first tested the hybridization assay in blocks of Calu-6 cell pellets (**Figure 7A**). The ISH assay showed abundant *LAPTM4B* mRNA expression in Calu-6 cells

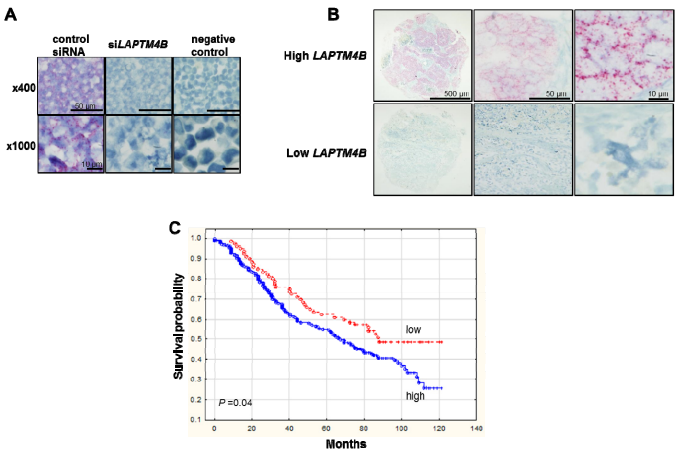

Figure 7. Analysis of *LAPTM4B* mRNA expression in lung cancer histological tissue specimens. A. Assessment of *LAPTM4B* mRNA by ISH was performed in blocks ofCalu-6 cell pellets transfected with control or *LAPTM4B*-specific siRNA. **B.** ISH was performed on a tissue microarray of 459 NSCLC FFPE specimens. C. Survival analysis based on *LAPTM4B* expression was performed using the Kaplan-Meier method for survival probability and the log-rank test.

transfected with control siRNA (**Figure 7A** , left) compared to cells transfected with *LAPTM4B*-specific siRNA (**Figure 7A**, middle) and no reactivity in cells after omitting the probe (**Figure 7A**, right). These data demonstrate reliability of this assay to detect specific *LAPTM4B* expression and at variable levels. We then applied the ISH assay to study *LAPTM4B* expression in NSCLC FFPE histological tissue specimens (303 adenocarcinomas and 156 SCCs) and found that *LAPTM4B* expression was confined to epithelial tumor cells and absent in the stroma (**Figure 7B**). *LAPTM4B* expression was then quantified and statistically analyzed in association with various clinicopathological information including clinical outcome. We found that LAPTM4B mRNA by ISH was significantly higher in males, older patients and notable in smokers (all P <0.05). Additionally, higher (greater than the median) *LAPTM4B* mRNA expression was significantly associated with worse overall survival (P <0.05 of the log-rank test) in comparison to lower *LAPTM4B* mRNA in lung adenocarcinoma patients (n=303) (**Figure 7C**). These findings suggest that *LAPTM4B* field cancerization marker is associated with smoking and poor clinical outcome in the pathogenesis of human lung cancer.

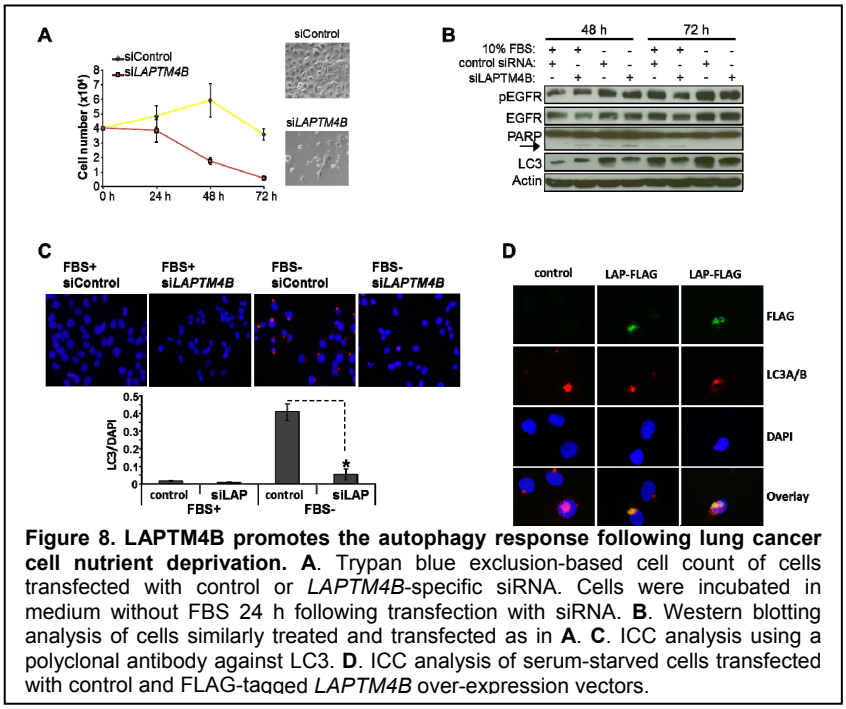

Figure 8. LAPTM4B promotes the autophagy response following lung cancer cell nutrient deprivation. A. Trypan blue exclusion-based cell count of cells transfected with control or *LAPTM4B*-specific siRNA. Cells were incubated in medium without FBS 24 h following transfection with siRNA. **B**. Western blotting analysis of cells similarly treated and transfected as in **A**. **C**. ICC analysis using a polyclonal antibody against LC3. **D**. ICC analysis of serum-starved cells transfected with control and FLAG-tagged *LAPTM4B* over-expression vectors.

We then determined to further examine the role of *LAPTM4B* field cancerization marker and putative oncogene in lung cancer pathogenesis. Earlier reports by others have demonstrated that *LAPTM4B* mediates breast cancer cell survival following metabolic and genotoxic stress [10-12]. We were prompted to examine the relevance of *LAPTM4B* expression to effects of nutrient deprivation in lung cancer cells. We compared and contrasted the effects of serum starvation on cells transfected with control and *LAPTM4B*-specific siRNA. RNA interference-mediated knockdown of *LAPTM4B* significantly and largely augmented cell growth inhibition induced by serum starvation (**Figure 8A**). In addition, western blotting analysis demonstrated that knockdown of *LAPTM4B* alone in basal conditions decreased phosphorylation of the epidermal growth factor receptor (EGFR) proto-oncogene (**Figure 8B**). Additionally, knockdown of

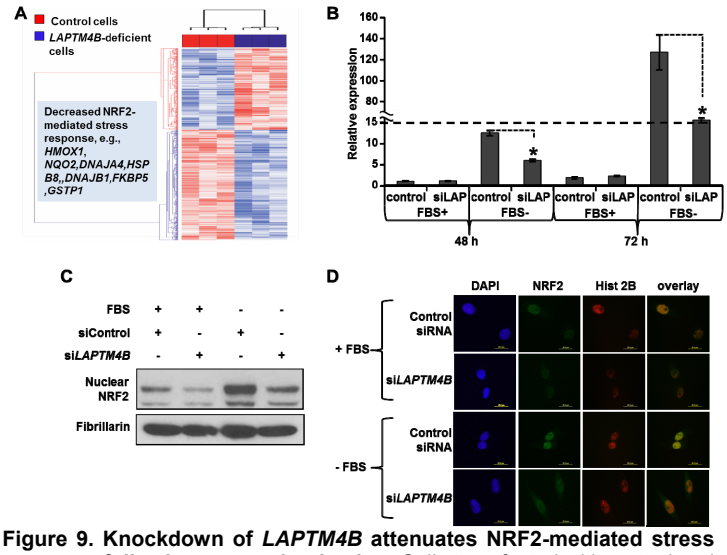

Figure 9. Knockdown of *LAPTM4B* attenuates NRF2-mediated stress response following serum deprivation. Cells transfected with control and *LAPTM4B*-specific siRNA and in the absence and presence of FBS were analyzed by global expression profiling (**A**), QRTPCR for *HMOX1* expression (**B**), western blotting for assessment of NRF2 levels in nuclear fractions (**C**) and by ICC to determine co-localization of NRF2 with the nuclear marker histone 2B (**D**). Error bars represent standard deviation. *P <0.05.

LAPTM4B increased serum starvation-induced cleavage of poly (ADP) ribose polymerase (PARP), indicative of augmented apoptosis induction. Notably, knockdown of *LAPTM4B* expression nearly abrogated the induction of the autophagy marker, LC3 [10], by serum starvation (**Figure 8B**). We then determined to quantify the levels of LC3 protein by immunocytochemical (ICC) analysis. Knockdown of *LAPTM4B* significantly attenuated (8-fold) LC3 induction by 48 h of serum starvation (P <0.001) (**Figure 8C**). We then performed ICC in serum-starved cells transfected with control vectors as well as a vector over-expressing a FLAG-tagged *LAPTM4B* construct. This analysis demonstrated that following serum starvation, LAPTM4B protein and LC3 co-localize intracellularly (**Figure 8D**). It is worthwhile to note that these findings are in accordance with earlier reports by others that pointed to the crucial role of *LAPTM4B* in the stability of the autophagosome in breast cancer cells [10-11].

We then determined to further understand the mechanisms that underlie augmented growth inhibition of lung cancer cells by decreased or absent *LAPTM4B* expression following serum deprivation. We performed global expression profiling, using the Human Gene 1.0 ST platform (Affymetrix), to compare and contrast the effect of serum starvation on the transcriptome of cells transfected with control and *LAPTM4B*-specific siRNA (**Figure 9A**). Expression data were normalized by RMA [13] and log (base 2) transformed. Gene features differentially modulated in the presence and absence of FBS between cells transfected with control and *LAPTM4B*-specific siRNA were determined by a mixed-effects model and a false discovery rate (FDR) cut-off of 5%. Ratios for the levels of each differentially expressed gene in the absence of serum over those in the presence of serum were determined. The ratios were then analyzed by clustering among and control and *LAPTM4B* deficient cells (**Figure 9A)**. Functional pathways analysis predicted significantly decreased activation of the nuclear factor erythroid 2-like 2 (NFE2L2 also known as NRF2)-mediated response [14] following serum withdrawal in *LAPTM4B* deficient cells (**Figure 9A**). We then performed QRTPCR analysis of heme oxygenase 1 (*HMOX1*) which is known to be transcriptionally modulated by *NRF2* [15] and was found by our array analysis to be decreased following knockdown of *LAPTM4B* (**Figure 9A**). QRTPCR analysis demonstrated that *HMOX1* mRNA levels were significantly increased after 48 h and 72 h of serum withdrawal (**Figure 9B**). Knockdown of *LAPTM4B* significantly attenuated *HMOX1* induction by serum withdrawal (**Figure 9B**). We then sought to assess the effect of *LAPTM4B* expression on the NRF2 transcription factor itself. In accordance with QRTPCR analysis of its downstream target *HMOX1*, knockdown of *LAPTM4B* attenuated nuclear accumulation of NRF2 protein by serum starvation as evident by western blotting of nuclear cell fractions (**Figure 9C**) and by ICC analysis of the co-localization of NRF2 with the nuclear marker histone 2B (**Figure 9D**). Furthermore, reciprocal effects were observed in cells transfected with *LAPTM4B*-overexpressing vectors. Over-expression of *LAPTM4B* significantly augmented *HMOX1* induction by serum starvation (**Figure 10A**) as well as the nuclear accumulation of NRF2 (**Figure 10B**). It is noteworthy that while modulation of *LAPTM4B* expression did not affect levels of *HMOX1* in cells at basal conditions and cultured in FBS-containing medium (**Figures 9B and 10A**), *LAPTM4B* expression positively controlled nuclear levels of the transcription factor NRF2 (**Figures 9C and 10B**). It is reasonable to surmise that *HMOX1* induction by NRF2 is independent on the levels of the latter transcription factor but rather dependent on serum deprivation-induced stress. Our findings point to a novel intracellular mechanism, which involves the *LAPTM4B* field cancerization marker, for control of the NRF2 transcription factor during basal conditions and cellular stress (e.g. nutrient deprivation).

Studies exploring mechanisms of the oncogenic properties and function (**Figures 7-10**) of *LAPTM4B* field cancerization marker

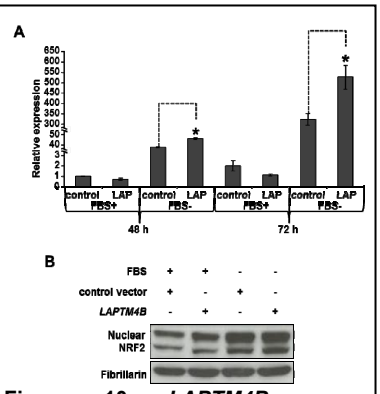

Figure 10. *LAPTM4B* over-expression augments *HMOX1* induction by serum starvation as well as NRF2 nuclear levels. A. QRTPCR analysis of *HMOX1* in cells transfected control and *LAPTM4B* overexpressing vectors in the presence and absence of FBS. **B**. Western blotting analysis of NRF2 levels in nuclear fractions. Membranes were incubated with antibody against the nucleolar marker fibrillarin to ensure equal loading. Error bars represent standard deviation. *, P <0.05.

will be completed and then be prepared as a manuscript for publication which we anticipate submitting for peer-review within the next funding period.

E. Collection and transcriptome analysis of nasal and bronchial epithelia from patients with and without lung cancer (Sub-specific Aims 1A and 1C)

As detailed in our previous annual report (Year 2), epithelial brushings were being collected from the nasal compartment and from the airway (3 bronchial brushes). Tumor, normal lung tissue and airway samples were also collected from specimens resected following surgical lobectomy procedures. Lung tumors, normal lung tissue and airway samples from cases with lung cancer were obtained from all four participating institutions and samples from cases without lung cancer were collected and processed from BU (Partnering PI, Dr. Avrum Spira), Vanderbilt (Partnering PI, Dr. Pierre Massion) and UCLA (Initiating PI, Dr. Steve Dubinett). Total RNA from all samples in the different institutions were isolated similarly using the miRNeasy kit from Qiagen according to the manufacturer's instructions.

During the past year (Year 3), we started studying the molecular spatial map of field effects that transverse the normal-appearing bronchus adjacent to tumors up to the relatively distant nasal epithelium (**Figure 11A**). We surmised that this analysis would aid in identification of shared genomic changes between the field and lung cancer and that extend to compartments (e.g. nasal) in the field cancerization that can be readily accessible for biomarker analysis in screening and clinical settings. Samples (n=254) from patients with (n=28) and without (n=9) lung cancer that were collected from all partnering institutions (14 cancer cases from MD Anderson) were processed for global expression profiling at MD Anderson Cancer center using the Human Gene 2.0 ST platform (Affymetrix) and data analysis was performed in collaboration with BU (Partnering PI, Dr. Avrum Spira). We identified profiles differentially expressed between NSCLCs and benign nodules which were then analyzed by GSEA. The top 50 gene sets significantly (P <0.01) enriched in the NSCLC tumors were then analyzed and plotted across different spatial points in the field of injury/cancerization. Spatial analysis of molecular field profiles pointed to gene sets that decreased in tumor-associated enrichment with increasing distance of samples from tumors and those (highlighted in red) that remain enriched up to the nasal epithelium (**Figure 11B**). These data point to specific tumor-associated profiles that are enriched in the nasal epithelium and, thus, comprise readily accessible field cancerization markers for lung cancer detection. We are currently completing the microarray data analysis and subsequent validation studies which we expect to prepare as a manuscript and submit for publication within the next funding period.

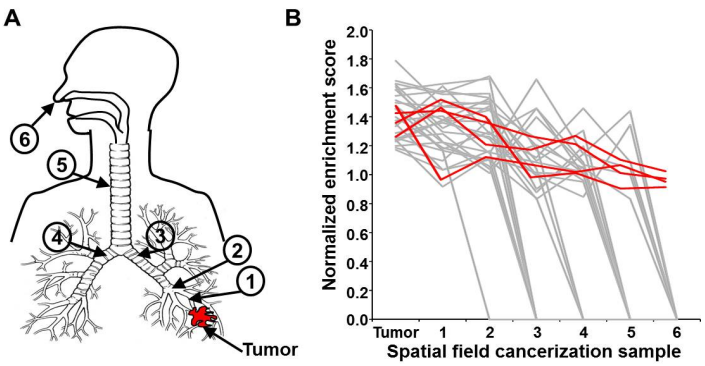

Figure 11. Spatial mapping of the field cancerization in NSCLC. Profiles in lesions (NSCLC tumors and benign nodules), airways adjacent to lesions (1 and 2), ipsilateral (3) and contralateral (4) main stem bronchi, tracheas (5) and nasal epithelia (6) were compared between NSCLC patients and those with benign disease (**A**). **B**. Profiles that were differentially expressed between airways 1 and 2 from patients with and without lung cancer were then analyzed by GSEA. The top 50 positively enriched (in cancer patient) pathways were then plotted for their enrichment across different spatial locations (1-6) in the field of cancerization. Plots indicate NSCLC-associated gene sets that decrease in positive enrichment in samples with increasing distance from tumors (grey) or remain highly enriched up the nasal epithelium (red).

Specific Aim 2: Evaluate the role of airway epithelium tumor-initiating stem/progenitor cells in current and former smokers.

Summary of Research Findings:

A. Assessment of the molecular profiles of tumor-initiating stem/progenitor cells from normal airway epithelium, premalignant lesions and cancer.

During the third year of funding, the UCLA team in collaboration with the BU team performed studies proposed in Aim 2, including RNA sequencing on laser-microdissected representative cell populations along the SCC pathological continuum of patient-matched normal basal cells, premalignant lesions, and tumor cells. We discovered transcriptomic changes and identified genomic pathways altered with initiation and progression of SCC within individual patients.

Fresh frozen tissue blocks were obtained from four individuals (patients 1-4) with lung SCC at the time of tumor resection, and regions of normal basal cells (BC), premalignant (squamous metaplastic and dysplastic) cells, and tumor cells were captured from sectioned tissues by laser microdissection. Sequencing libraries of the expected concentration and cDNA size ranges were generated from RNA isolated from the microdissected cells. All sequenced samples produced reads with mean Phred quality scores above 25, indicating that it was possible to generate sequencing libraries of good quality with our method of isolating RNA from laser-microdissected materials. Because of this large amount of variability, reads aligning to the mitochondrial genome were discarded from analysis after alignment, and RPKM (reads per kilobase per millions of reads) values were computed relative to the total number of reads aligning uniquely to the nuclear genome. To identify SCC-associated genes whose expression is also associated with progression from normal airway BC to premalignant (metaplastic or dysplastic) lesions, a multi-step procedure as used as outlined in **Fig. 12A**. We identified 626 early-stage genes (significantly differentially expressed in a similar manner in both premalignant lesions and tumor compared to normal BC), 730 late-stage genes (significantly differentially expressed in a similar manner in tumor compared to both premalignant lesions and normal BC), and 68 "stepwise" genes (significantly differentially expressed in both of the described stages of carcinogenesis) (**Fig. 12B**). We selected three genes for further validation: *CEACAM5, SLC2A1* and *PTBP3*. These genes, whose expression was upregulated in premalignant lesions and tumor cells compared to normal BC, were chosen because of their potential roles in the biology of lung carcinogenesis. The expression of *CEACAM5* and *SLC2A1* was measured by performing qPCR on remaining material from the sequencing libraries of patients 3 and 4, as well as on laser-microdissected RNA from four additional independent cases (patients 5-8). In each case, the mRNA level of each gene was significantly higher (sign test $p < 0.05$) in the premalignant lesion than in normal BC (**Fig. 13A**). Because mRNA and protein levels may not always be well correlated, immunofluorescent staining was performed in sections of normal epithelium, premalignant lesion, and carcinoma from two independent cases (patients 9 & 10). CEACAM5 (**Fig. 13B**) and SLC2A1 (not shown) were not detectable in the normal epithelia, but they were highly expressed in cells within both metaplastic lesions and the SCC tumors. SLC2A1 was expressed throughout the KRT5$^+$ component of the tumor, whereas CEACAM5 was expressed in some, but not all, KRT5$^+$ tumor cells. PTBP3 was strongly expressed in premalignant lesions and tumor cells, and although it was strongly expressed in

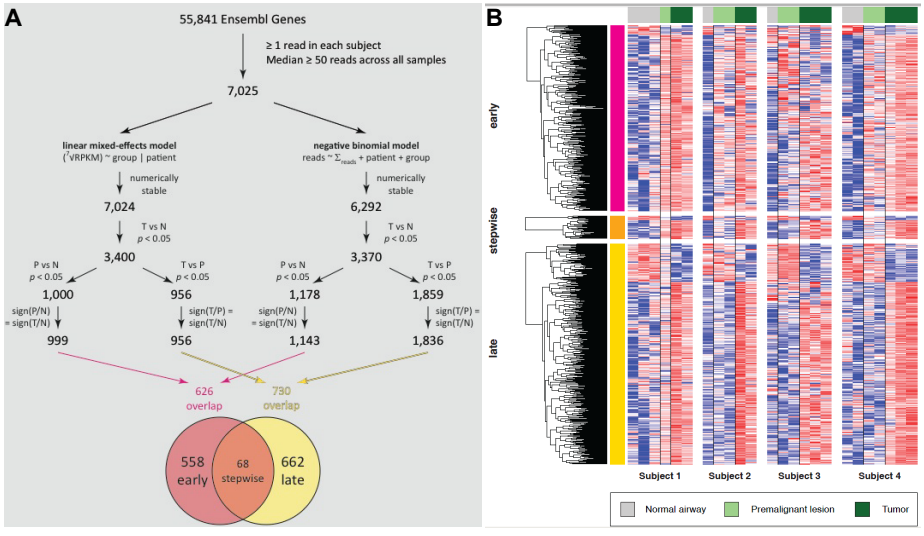

Figure 12. Identification of genes associated with early- or late-stage SCC carcinogenesis. **A**. Analysis flowchart. Uniquely aligned reads were assigned to 55,841 Ensembl Gene loci (Ensembl build 69). Genes with significant (p<0.05) differential expression between tumor and normal cells, as well as between premalignant and normal cells ("early" genes), between tumor and premalignant cells ("late" genes), or both ("stepwise" genes) were identified. **B**. Expression heatmap. Red and blue indicate genes with expression that is higher or lower than the mean within each patient, respectively. Genes are hierarchically clustered within each group (early, stepwise, late).

columnar KRT5$^-$ cells of normal airway epithelium, its expression was undetectable in normal BC. To better understand the biological role these genes may play in the development of lung SCC, the Gene Expression Omnibus (GEO) Profiles tool was used to examine their expression in GEO DataSets associated with experimental parameters relevant to lung SCC carcinogenesis. First, *SLC2A1* and *PTBP3* were confirmed to be significantly upregulated in an independent set of SCC tumors with respect to paired samples of adjacent normal tissue; however, the expression of *CEACAM5* was unchanged. Next, a collection of SCC and adenocarcinoma (ADC) lung tumors (GDS3627) was interrogated to determine the specificity of the expression of these genes with respect to the SCC tumor type. The expression of *SLC2A1* and *PTBP3* were again strongly increased in SCC tumors compared with ADC tumors; however, *CEACAM5* was moderately downregulated in SCC relative to ADC. Finally, because premalignant lesions in large central airways are believed to arise from injury caused by cigarette smoking, the expression levels of these genes were examined in a study of bronchoscopic brushings of healthy current, former, and never smokers (GDS534) [5]. In this study, *CEACAM5* and *SLC2A1* were significantly upregulated in brushings from current smokers compared with those from never smokers, although *PTBP3* was not.

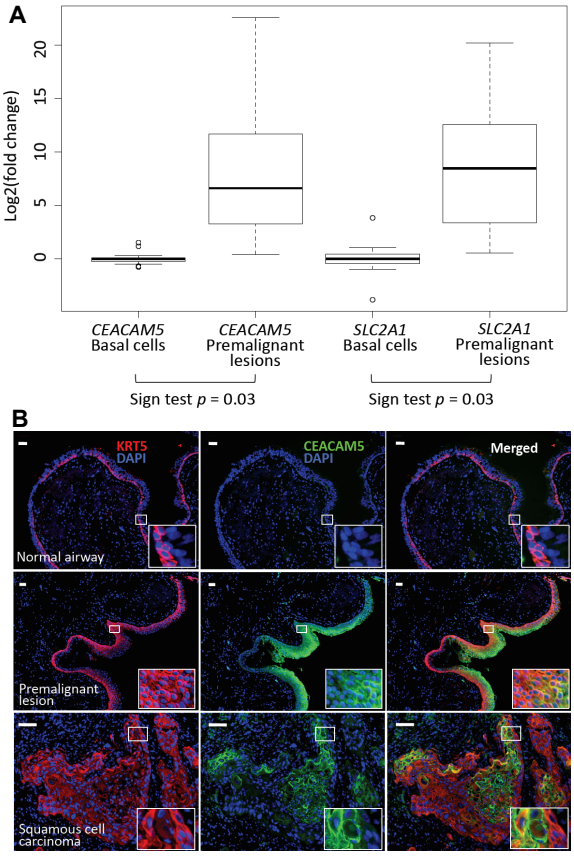

Figure 13. Experimental validation of *CEACAM5* expressi-on. **A**. Quantitative RT-PCR. Box plots represent RNA levels of *CEACAM5* and *SLC2A1* in normal BC and premalignant lesions from six patients. **B**. Immunofluorescent staining of CEACAM5. Protein staining shows increased expression of CEACAM5 in premalignant lesions and SCC compared to BC in the normal epithelium. Left columns: KRT5 (red), marker for BC, premalignant lesions, and tumor cells; middle columns: CEACAM5 (green); right columns: merged images of left and middle columns. DAPI (blue), nuclear marker.

B. Prediction of chromosomal gains and losses during carcinogenesis

Gene Set Enrichment Analysis (GSEA) performed using positionally defined gene sets (cytobands) revealed that late-stage (but not early-stage) carcinogenesis is associated with a coordinate loss of expression in the p arm of chromosome 3 and an attendant gain of expression in 3q26.33-3q29 (**Fig. 14A**), which corresponds to previously reported observations of frequent 3p deletion and 3q amplification in squamous tumors [16-17]. In particular, chromosomal band 3q26.33 has been reported to be consistently amplified in lung SCC [18].

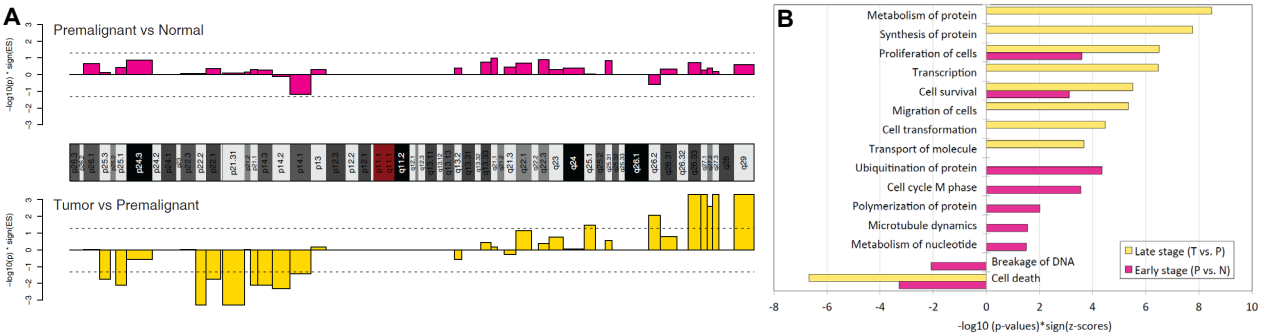

Figure 14. Identification of coordinately regulated chromosomal regions and pathways. **A**. Identification of differenti-ally regulated cytobands by GSEA. Dashed lines indicate nominal p=0.05. **B**. Identification of dysregulated biologi-cal functions by IPA. Selected biological functions (p<0.05 and z-scores ≥2 or ≤-2) predicted to be significantly increased (positive X axis values) or decreased (negative X axis values) in early-stage (magenta bars) and late-stage (yellow bars) carcinogenesis.

C. Identification of biological changes in early- and late-stage carcinogenesis

Ingenuity Pathway Analysis (IPA) (Ingenuity Systems) was used to further characterize the changes in biological functions resulting from the differential expression of genes associated with early-stage events, which contribute to the initiation and formation of premalignant lesions, or with late-stage events, which are involved in the progression from premalignant lesions to tumor. This analysis revealed that the early-stage carcinogenesis was characterized uniquely by increased protein ubiquitination and cell cycle progression, whereas the late-stage events were marked primarily by increased transcriptional and translational activity and cellular migration and transformation (**Fig. 14B**). In addition, an increase in cell survival and proliferation and a corresponding downregulation of cell death mechanisms was observed throughout both stages of carcinogenesis. IPA was also used to determine whether the genes identified to be differentially expressed either early or late in carcinogenesis are enriched in known targets of various transcription factors. This approach revealed that the set of genes that is differentially expressed early in carcinogenesis and remains dysregulated in tumor cells is enriched in previously reported targets of MYC and TP53 (**Fig. 15A**). To test the hypothesis that MYC activity is induced during early SCC carcinogenesis, immunofluorescent staining of MYC was performed to examine its nuclear and cytoplasmic localization in normal BC, premalignant lesions, and tumor cells (**Fig. 15B**). MYC staining was exclusive to the nuclei of premalignant lesions and tumor cells. In the histologically normal BC of the airways from patients with lung cancer, however, MYC was localized predominantly in the cytoplasm, although some areas of nuclear staining were also seen. The increased expression of MYC targets in the premalignant lesions and tumor cells, together with a concomitant increase in the nuclear localization of MYC, is strong evidence for a carcinogenesis-associated increase in MYC activity without a significant increase in gene expression.

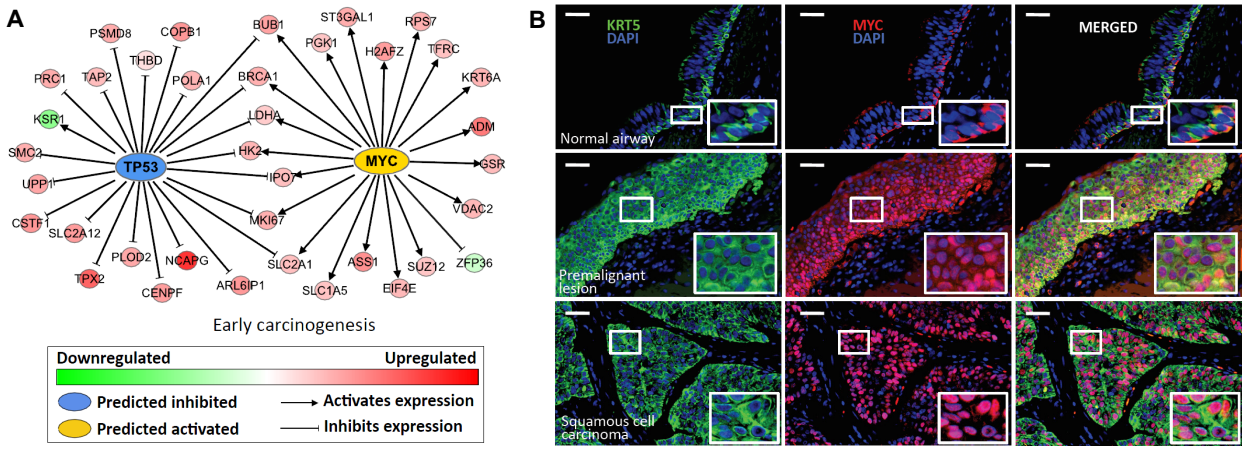

Figure 15. Identification of MYC as a dysregulated transcription factor in early carcinogenesis. **A**. Identification of dysregulated transcription factors by IPA. Expression patterns of known downstream targets of TP53 and MYC suggest the inhibition of TP53 and the activation of MYC during early carcinogenesis. **B**. Experimental validation of MYC activation in premalignant lesions and SCC. Immunofluorescent staining shows detection of MYC expression in the nuclei of premalignant lesions and SCC compared to cytoplasmic localization in BC.

D. Investigating molecular signature of lung cancer development in bronchial specimens by reverse phase protein array (RPPA)

Table 2. Specimen type collected.

DOD pt. case	MRN	Patient ID	LCB ID	Sample Type	Fixative	Sample ID	Storage Temperature in Centigrade	Location of Brushes
Vanderbilt 11	33567728	8841	2012-4-1-800-1	Normal Tissue	RNA Later	534048	-80	Tumor is central - LUL
				Normal Tissue		534049	-80	B1 closest to tumor
				Tumor Tissue	RNA Later	534050	-80	B2 same bronchus as B1 - peripheral
				Tumor Tissue		534051	-80	B3 and B4 - different airway -
				Brushes B1-B4	Qiazol	534052-55	-80	
Vanderbilt 12	34227843	8836	2012-4-1-802-1	Normal Tissue	RNA Later	534293	4	Tumor is peripheral (s/p chemorad) - RUL
				Normal Tissue		534294	-80	B3 closest to tumor
				Tumor Tissue	RNA Later	534295	4	B2 same airway as B3 - more proximal
				Tumor Tissue		534296	-80	B1 different airway - proximal
				Brushes B1-B3	Qiazol	534299-301	-80	
								scant tumor left on specimen after chemorad
Vanderbilt 13	34026005	8844	2012-4-1-803-1	Normal Tissue	RNA Later	534326	4	Tumor is central - LUL
				Normal Tissue		534327	-80	B1 closest to tumor
				Tumor Tissue	RNA Later	534329	4	B2 same airway as B1 - distal
				Tumor Tissue		534328	-80	B3 different airway
				Brushes B1-B3	Qiazol	534330-32	-80	
Vanderbilt 14	31374218	9002	2012-5-1-811-1	Normal Tissue	RNA Later	535568	4	Tumor is peripheral - RLL
				Normal Tissue		535569	-80	B4 closest to tumor
				Tumor Tissue	RNA Later	535570	4	B1 and B2 same airway as B4 - more proximal
				Tumor Tissue		535571	-80	B3 different airway - proximal
				Brushes B1-B4	Qiazol	535572-75	-80	
Vanderbilt 15	33737560	9006	2012-5-1-814-1	Normal Tissue	RNA Later	535692	4	Tumor is central - LUL
				Normal Tissue		535690	-80	B1 closest to tumor
				Tumor Tissue	RNA Later	535693	4	B2 and B4 same airway as B1 - distal
				Tumor Tissue		535691	-80	B3 different airway
				Brushes B1-B4	Qiazol	535694-97	-80	
Vanderbilt 16	8993586	9047	2012-6-1-821-1	Normal Tissue	RNA Later	536361	4	Tumor is Central - LUL
				Normal Tissue		536360	-80	B1 is closest to tumor
				Tumor Tissue	RNA Later	536363	4	B2 on a different airway than B1 distal.
				Tumor Tissue		536362	-80	B3 on an opposite airway distal.
				Brushes B1-B3	Qiazol	536364-66	-80	
Vanderbilt 17	15991904	9078	2012-6-1-824-1	Normal Tissue	RNA Later	536782	4	Tumor is peripheral - LLL
				Normal Tissue		536781	-80	B1 is distal to tumor
				Tumor Tissue	RNA Later	536784	4	B2 is closest to tumor on same airway as B1.
				Tumor Tissue		536783	-80	B3 is on a different airway distal.
				Brushes B1-B3	Qiazol	536785-87	-80	
Vanderbilt 18	32690034	9138	2012-7-1-828-1	Normal Tissue	RNA Later	537331	4	Tumor is peripheral - LUL
				Normal Tissue		537330	-80	B1 is closest to tumor.
				Tumor Tissue	RNA Later	537329	4	B2 on same airway as B1 distal.
				Tumor Tissue		537328	-80	B3 is on a different airway distal.
				Brushes B1-B3	Qiazol	537332-34	-80	
Vanderbilt 19	34594325	9258	2012-8-1-840-1	Normal Tissue	RNA Later	538648	4	Tumor is perripheral - LLL
				Normal Tissue		538649	-80	B2 is closest to tumor on the same airway as B1.
				Tumor Tissue	RNA Later	538646	4	B3 and B4 are distal to tumor on different airway.
				Tumor Tissue		538647	-80	
				Brushes B1-B4	Qiazol	538650-653	-80	
Vanderbilt 20	3467849	9341	2012-9-1-842-1	Normal Tissue	RNA Later	539306	4	Tumor is peripheral - RUL
				Normal Tissue		539307	-80	B1 is distal to tumor.
				Tumor Tissue	RNA Later	539304	4	B2 is distal on same airway as B1.
				Tumor Tissue		539305	-80	B3 is on a different airway distal.
				Brushes B1-B4	Qiazol	539308-311	-80	B4 is proximal to the tumor on same airwaiy as B2 and B1.
Vanderbilt 21	23268162	9401	2012-9-1-848-1	Normal Tissue	RNA Later	539677	4	Tumor is peripheral - RUL
				Normal Tissue		539678	-80	B1 is proximal to tumor on same airway as B2.
				Tumor Tissue	RNA Later	539675	4	B2 is distal to tumor.
				Tumor Tissue		539676	-80	B3 is distal to tumor on different airway.
				Brushes B1-B3	Qiazol	539679-681	-80	
Vanderbilt 22	19271741	9519	2012-11-1-862-1	Normal Tissue	RNA Later	541376	4	Tumor is peripheral -RUL
				Normal Tissue		541377	-80	B2 is proximal to tumor on same airway as B1.
				Tumor Tissue	RNA Later	541378	4	B1 is distal to tumor.
				Tumor Tissue		541379	-80	B3 is distal to tumor on different airway.
				Brushes B1-B3	Qiazol	541380-382	-80	
Vanderbilt 23	35129154	9659	2013-1-1-876-1	Normal Tissue	RNA Later	542448	4	Tumor is peripheral - RUL
				Normal Tissue		542449	-80	B1 is closest to tumor.
				Tumor Tissue	RNA Later	542451	4	B2 is distal to tumor on same airway as B1.
				Tumor Tissue		542450	-80	B3 is distal to tumor on different airway.
				Brushes B1-B4	Qiazol	542452-454	-80	
Vanderbilt 24	35655885	9979	2013-4-1-896-1	Normal Tissue	RNA Later	544671	4	Tumor is peripheral - RUL
				Normal Tissue		544672	-80	B1 is proximal to tumor, about .5cm from tumor.
				Tumor Tissue	RNA Later	544673	4	B2 is distal to tumor, about 2.5 cm from tumor. B2 and B1 on same airway.
				Tumor Tissue		544674	-80	B3 is distal, about 3cm from tumor and is on different airway.
				Brushes B1-B3	Qiazol	544675-677	-80	
Vanderbilt 25	29961091	10477	2013-9-1-917-1	Tumor Tissue		546981	-80	
				Tumor Tissue	RNA Later	546982	4	
				Normal Tissue		546983	-80	Tumor is central - Left pnemonectomy
				Normal Tissue	RNA Later	546984	4	B1 is distal to tumor
				Brushes B1-B3	Qiazol	546985-987	-80	B2 is closest to tumor. B2 and B3 are in the same airway.

We investigated the expression of a group of selected proteins in bronchial brushings and biopsies from low risk and high risk individuals without lung cancer and lung cancer patients by reverse phase protein array (RPPA). Risk level of individuals without lung cancer was calculated based on Bach model criteria (age, gender, smoking history and asbestos exposure) [19]. The goal is to identify proteins and pathways altered by risk factors in the field of cancerization. The hypothesis is that a subset of functionally significant molecular alterations in the airway epithelium represents early steps in tumorigenesis, and its measure in individuals at risk for lung

cancer development may allow us to derive new insights into tumorigenesis and a tool for risk assessment that will provide the basis of patient selection for surveillance programs.

Methods

We conducted two experiments referred as experiment 1 and experiment 2 using two independent sets of airway specimens. The first experiment was performed using pairs of bronchial brushings and biopsy specimens from 15 patients, 5 per group (low risk, high risk, and cancer group). Brushings and biopsies were collected from the same individual at the same time. The second experiment was performed using biopsies only, each group consisting of 10 biopsies. All experiments were conducted under IRB approved protocol.

Specimen collection procedure: Bronchial brushings were collected from patients under conscious sedation. The brushings were immediately dipped into 1.5 ml saline taken in a labeled eppendorf tube. The tube was kept on ice to minimize protease action. Care was taken to keep the brush specimen free from blood. Brushings in the saline was vigorously agitated by vortexing for about 10 seconds with highest speed. It was then spun 1500g for 10 minutes in a microcentrifuge with the brush inside the tube. Supernatant was removed carefully leaving as little saline as possible keeping the brush inside the tube. The pellet was stored in freezer at -80^{o}C temperature. Patients undergoing autofluorescence bronchoscopy for clinical suspicion of lung cancer agreed to provide bronchial biopsy specimens at predetermined normal sites (with normal fluorescence ratio). Biopsy specimens collected for research were snap frozen and stored in -80^{o}C freezer. RPPA was performed according to the published protocol [20].

Statistical Methods: ANOVA or two sample t test is applied on a marker-by-marker base to test whether there is difference between the two sources within three sample groups. ANOVA and Tukey's HSD testis applied on a marker-by-marker base to test whether there are differences among the three risk groups and with comparisons contribute to the difference. Because of the multiple testing involvedin this approach, the individual ANOVA p-values are not particularly meaningful. However, when we look across the entire set of tests, the distribution of the p-values (under the null hypothesis that no RPPAs provide useful information) should be uniform. If, on the other hand, some RPPAs provide useful information about predicting the response, we would expect an overabundance of small p-values. We can capture this situation by modeling the distribution of the p-values with a Beta-uniform Mixture (BUM). To identify significantly differentially expressed RPPAs, we choose a cutoff for the single test p-values by controlling the false discovery rate (FDR), which is defined as the percentage of RPPAs called significant that are expected.

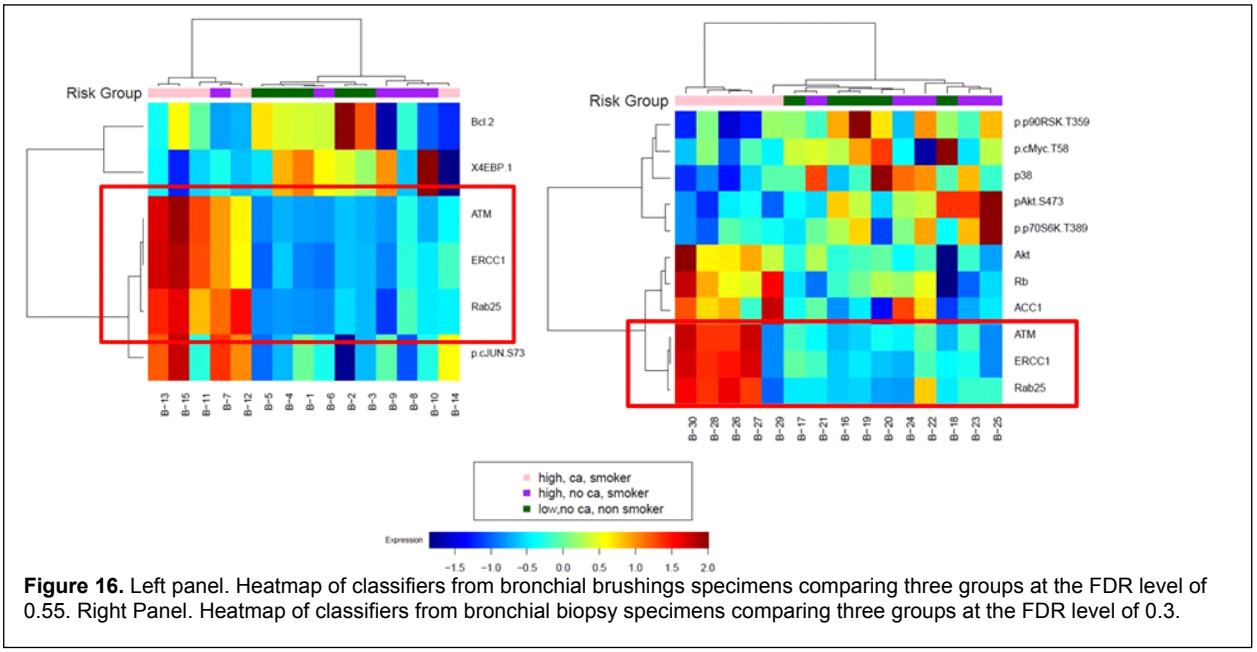

Figure 16. Left panel. Heatmap of classifiers from bronchial brushings specimens comparing three groups at the FDR level of 0.55. Right Panel. Heatmap of classifiers from bronchial biopsy specimens comparing three groups at the FDR level of 0.3.

Results

Hierarchical cluster analysis of the results from first experiment was performed with bronchial brushings and bronchial biopsy specimens separately. Applying linear regression model we checked whether the RPPA data show different expression patterns for the three groups. Heatmaps were generated using RPPA data from brushings and biopsy specimens comparing three groups at FDR level of 0.55 and 0.3 respectively (**Figure 16**). Four out of five brushings of the cancer group was clustered together (**Figure 16**. Left panel) and all five biopsies from cancer group were grouped together **Figure 16**. Right panel). Among the classifiers of both type of specimens, ATM, ERCC1 and Rab25 demonstrated overexpression in specimens from cancer patients.

The second experiment was performed with an independent set of bronchial biopsy specimens. Bronchial brushings were not used in this experiment. Comparison of three groups resulted two main clusters as sown in **Figure 17**. Left cluster includes all control biopsies, 2 out of 10 high risk biopsies and 4 out of 10 biopsies from cancer group. Right cluster is consisting of rest of the high risk and cancer biopsies. Upregulation of Notch3, EGFRpY1173, Axl, Mre11, LCN2 and BRCA1 in almost all samples in the left cluster and downregulation in the right cluster are nicely demonstrated. Unlike experiment 1, ERCC1 was upregulated only in 3 out of 10 biopsies of the cancer group. Rb.pS807.811 was overexpressed in the right cluster and underexpressed in the left cluster. ACC.pS79 was overexpressed in 3 biopsies of the cancer group.

Combining low risk and high risk individuals without cancer and comparing with patients with cancer of the second experiment resulted clusters that are not based on cancer status. Overall two clusters are noticeable mainly based on the expression of ATMpS1981 and ERCC1 in 3 out of 10 patients of the cancer group. Modest overexpression of these proteins was also observed in three low risk, one high risk and one cancer specimen. TSC2.pT1462 is overexpressed in patients without cancer (**Figure 18**. Left pane).

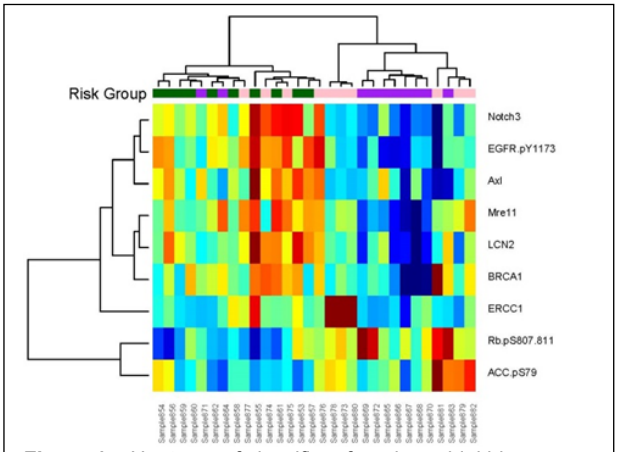

Figure 17. Heatmap of classifiers from bronchial biopsy specimens comparing three groups at the FDR level of 0.25.

Next, smokers without cancer and cancer patients who are also smokers were combined

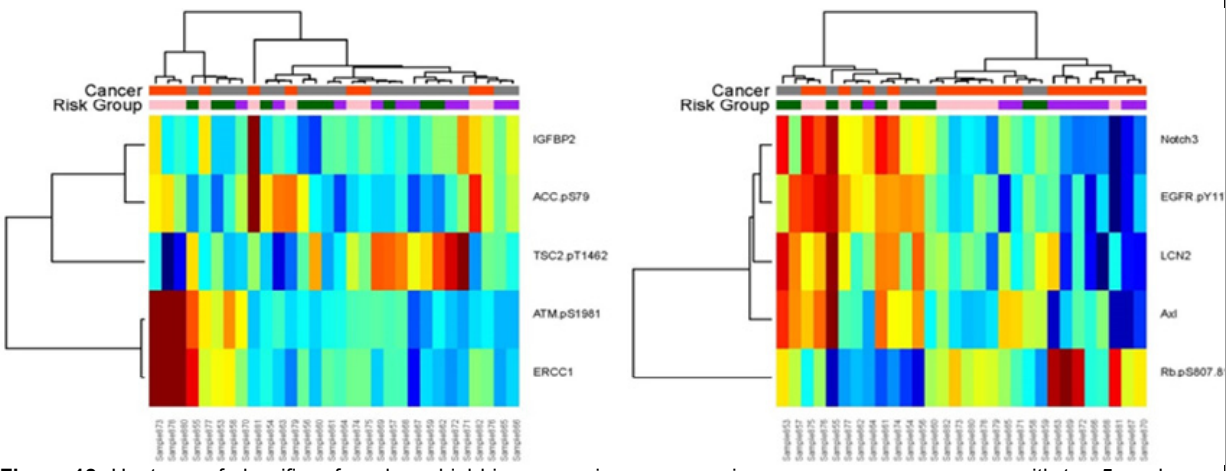

Figure 18. Heatmap of classifiers from bronchial biopsy specimens comparing cancer versus no cancer with top 5 markers, IGFBP2, ACCpS79, TSC2pT1462, ATMpS1981 and ERCC1 (left panel) and Heatmap of classifiers from bronchial biopsy specimencomparing non-smoker versus smoker with top 5 markers, Notch3, EGFR.pY1173, LCN2, AXL and Rb.pS807.811 (right panel).

and compared with low risk, non-smoking individuals, all from the second experiment. Two main clusters are identified in the heatmap of non-smoker versus smoker (**Figure 18**. Right panel). Two low risk clustered with high risk and cancer groups (right cluster). Notch 3, EGFRpY1173, LCN2, and Axl are downregulated in this cluster and overexpressed in 8 out of 10 nonsmokers (left cluster). Five out of 20 smokers (high risk and cancer groups) clustered with nonsmokers (low risk groups) where these proteins are overexpressed. Unlike above 4 markers Rb.pS807.811 was overexpression in 5 of the 20 smokers (right cluster) and downregulated in the other cluster.

The data from the second experiment was analyzed by separating the biopsies based on the subtypes of lung cancer found in the third group, i.e. adenocarcinoma (**Figure 19**. Left panel) and squamous cell (**Figure 19**. Right panel) carcinoma. There is indication of stronger classification of risk groups for squamous cell carcinoma subtype compared to adenocarcinoma.

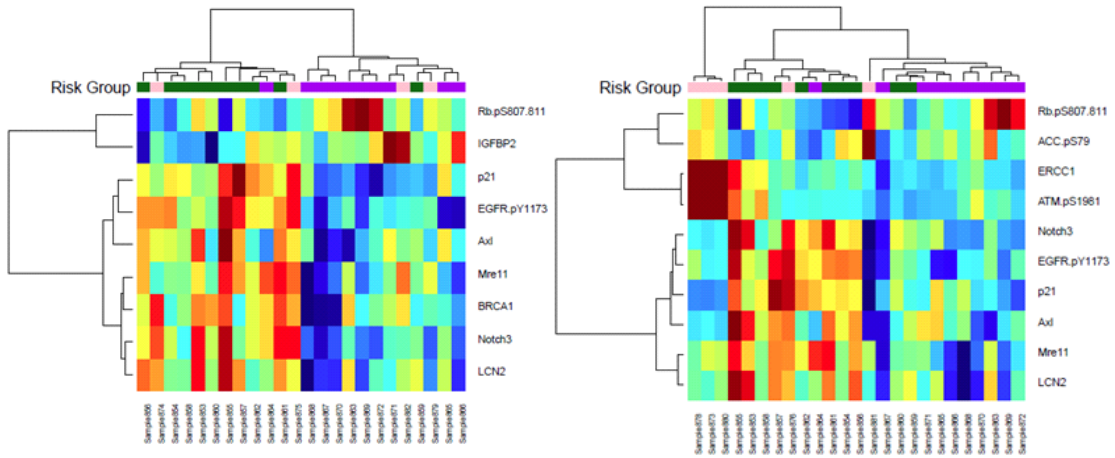

Figure 19. Heatmap of classifiers from bronchial biopsy specimens comparing low risk, high risk and adenocarcinoma only (left panel) and squamous cell carcinoma only (right panel) at the FDR levels of 0.55 and 0.15 respectively.

Future directions

In order to validate the results of the two experiments in a larger dataset we are currently acquiring biopsy specimens for the three groups described above. We will perform RPPA experiments with an expanded list of antibodies. In addition to the 140 proteins that include 125 proteins of the first experiment, we will investigate expression of metabolic enzymes because of possible metabolic reprograming of the airway epithelium, the field of cancerization, of at risk individuals which is indicated in the findings from another project in the laboratory. Expression of selected high ranking candidate proteins will be validated in specimens from collaborators at BU, UCLA and MDA.

In collaboration with the Dubinett laboratory, the Liebler laboratory is analyzing proteomic characteristics of Snail-driven malignant conversion in human bronchial epithelial cells (HBEC). Snail overexpression drives anchorage-independent growth (AIG). The goal of these analyses is to determine the requirements for P53, KRAS, or both P53/KRAS for full Snail-driven malignant conversion in vivo. Dr. Tonya Walser of the Dubinett laboratory has generated Snail over-expressing HBEC and HBEC-mutant NRH cell lines. Snail expression was confirmed by western blot and cells were confirmed as mycoplasma-negative. Genotyping confirmed P53 and KRAS mutation status and AIG assays were completed. Dr. Walser provided the following cell line pellets to the Liebler laboratory: HBEC2-Snail, HBEC2-vector control, HBEC11-Snail, HBEC11-vector control, H3mutP53/KRAS#12-Snail and H3mutP53/KRAS#12-vector control. Three independent replicate samples of each cell type are being analyzed on a standardized shotgun proteomics platform, in which cell proteins are digested with trypsin and fractionated by basic reverse phase chromatography. Fifteen concatenated fractions from each sample are then analyzed by reverse phase liquid chromatography-tandem mass spectrometry on a

Thermo Orbitrap Elite instrument. These analyses are in progress and will be completed by approximately October 31st.

Specific Aim 3: Test airway-based mRNA and microRNA biomarkers of diagnosing lung cancer in current and former smokers at high risk for lung cancer in minimally invasive sites.

Due to the use of both next generation RNA sequencing and comprehensive microarray profiling and due to this ongoing study's unique design we anticipate that expression profiles in the NSCLC molecular field of injury will harbor molecules, both novel and established, that may exhibit potential for use as airway biomarkers that can be developed and tested for lung cancer detection using minimally invasive sites in Specific Aim 3 of this award. As mentioned in Specific Aims 1 and 2 above, we have identified profiles in the field of injury/cancerization that are also enriched in the nasal compartment of patients with lung cancer relative to patients with benign disease. While microarray (MD Anderson and BU) and RNA sequencing (BU) data analyses are being completed, all four sites/institutions are continuing to collect nasal and airway samples from patients with lung cancer and BU, Vanderbilt and UCLA are continuing to collect nasal and airway samples from patients without lung cancer or benign disease. These new cases will serve as sets to develop and validate classifiers, that are based on profiles from Aim 1 as mentioned above, that can be analyzed readily in the clinic (e.g. by qRT-PCR) in minimally invasive sites (e.g. nasal compartment) in smokers with indeterminate nodules. In addition, to the fourteen lung cancer cases that have already been profiled in Specific Aim 1, as mentioned above, additional 43 lung cancer cases comprised of large airways, airways adjacent to the nearby lung tumor and nasal epithelia, have been collected at MD Anderson Cancer Center. These additional cases will be utilized for development of the classifier in Year 4 of the grant period.

Key Research Accomplishments

Aim 1

- Generated high quality Illumina mRNA seq and Affymetrix Gene ST 2.0 array data on airway epithelium collected throughout the field of injury using uniform SOPs at four institutions.
- Completed initial processing and QC of both datasets.
- Demonstrated related enrichment of cancer genes throughout multiple sites with the airway.
- Characterized the expression dynamics of the cancer field of injury in the airway and the transcriptomic architecture of the adjacent airway field cancerization in early-stage non-small cell lung cancer. These analyses demonstrated that the adjacent airway field of cancerization is comprised of markers that can identify lung cancer among smokers as well as gradient and localized site-dependent expression patterns that recapitulate NSCLC profiles. These findings have been submitted recently for publication and are under revision.
- Demonstrated for the first time that the field cancerization putative oncogene, *LAPTM4B*, is a positive mediator of the lung cancer cell malignant phenotype evidenced by its promotion of anchorage-independent colony formation in soft agar.
- Studied the mRNA expression of *LAPTM4B* in a large series of NSCLC histological tissue specimens for the first time by *in situ* hybridization. This analysis revealed that *LAPTM4B* expression is significantly positively associated with smoking and worse overall survival.
- Demonstrated that the field cancerization marker *LAPTM4B* protects lung cancer cells from serum deprivation-induced growth inhibition and promotes the autophagy response following serum deprivation.
- Revealed that *LAPTM4B* is a novel positive regulator of NRF2 transcription factor in lung cancer cells.
- In collaboration with the Partnering PIs and Initiating PI of this grant, performed microarray profiling at MD Anderson of 254 field cancerization samples from 28 cases with lung cancer and 9 cases with benign disease to begin to characterize the molecular spatial map of field effects that transverse the bronchus adjacent to tumors up to the nasal epithelium. This

Aim 2

- By profiling of gene expression in airway basal cells, premalignant lesions and tumors from the same patients, we identified coordinate changes in the activity of upstream regulators and the expression of downstream genes within the same patient during early- and late-stage carcinogenesis.
- Determined that expression of *CEACAM5* and *SLC2A1* genes was not detectable in the normal epithelia, but they were highly expressed in cells within both metaplastic lesions and the SCC tumors.
- Identified the transcription factors *TP53* and *MYC* as likely candidates based on the coordinate differential expression of their target genes.
- Demonstrated that late-stage (but not early-stage) carcinogenesis is associated with a coordinate loss of expression in the p arm of chromosome 3 and an attendant gain of expression in 3q26.33-3q29.
- Determined that bronchoscopy specimens like brushings and biopsy, can be used interchangeably for RPPA profiling.
- Differential protein expression was demonstrated to be the same in bronchial brushings and bronchial biopsy from the same patient by RPPA.
- Found that molecular alterations in the bronchial biopsyspecimens of at risk individuals that includes epithelium and submucosa could provide a signature for risk assessment.
- Discovered that among the differentially expressed candidate proteins ATMand ERCC1 are highly discriminatory among the groups suggesting DNA damage and repair activation in the high risk group.

Reportable Outcomes

Abstracts:

- Maki Y, Fujimoto J, Yoo SY, Gower A, Shen L, Garcia MM, Kabbout M, Chow CW, Hong WK, Kalhor N, Wang J, Moran C, Spira A, Coombes KR, Wistuba II, Kadara H. Transcriptomic architecture of the airway field cancerization in early-stage non-small cell lung cancer. 104th Annual American Association for Cancer Research (AACR) meeting, April 6 - April 10 2013, Washington, D.C. Abstract # 2367.
- Ooi AT, Gower AC, Zhang K, Vick J, Hong LS, Fishbein M, Nagao B, Wallace WD, Elashoff DA, Dubinett S, Lenburg M, Spira A, Gomperts BN. Gene expression alterations in premalignant lesions from the airways of patients with lung squamous cell carcinomas. Platform presentation and travel award. AACR Washington DC, April 2013.

Manuscripts:

- Ooi AT, Gower AC, Zhang KX, Vick JL, Hong L, Nagao B, Wallace WD, Elashoff DA, Walser TC, Dubinett SM, Pellegrini M, Lenburg ME, Spira A, Gomperts BN. Profiling premalignant lesions in lung squamous cell carcinomas identifies mechanisms involved in stepwise carcinogenesis. *Cancer Research*, submitted and under revision.
- Kadara H, Fujimoto J, Yoo SY, Maki Y, Gower AC, Kabbout M, Garcia MM, Chow CW, Chu Z, Mendoza G, Shen L, Kalhor N, Hong WK, Moran C, Wang J, Spira A, Coombes KR, Wistuba II. Transcriptomic architecture of the adjacent airway field cancerization in non-small cell lung cancer. *Journal of the National Cancer Institute*. Submitted and Under Revision.
- Perdomo C, Campbell JD, Gerrein J, Tellez C, Garrison CB,, Walser TC, Drizik E, Si H, Gower AC, Vick J, Anderlind C, Jackson JR, Mankus C, Schembri F, O'Hara C, Gomperts BN, Dubinett SM, Hayden P, Belinsky SA, Lenburg ME, Spira A. miR-4423 is a Primate-

Conclusion

During our third year of research, we have successfully generated high quality mRNA sequencing and mRNA array data. Linear modeling followed by gene set enrichment analysis reveal connections in cancer gene expression throughout the airway. Specifically, we found that cancer genes in the distal airway are enriched in genes up in cancer in the mainstem bronchus. Similarly, cancer genes enriched in the bronchus and main carina show moderate enrichment in the nose. In summary this study reveals spatially connected gene expression patterns in the airway of patients with lung cancer.

Also, during the current funding period we characterized the transcriptomic architecture of the adjacent airway field cancerization in early-stage NSCLC. Our studies demonstrated that the adjacent airway field of cancerization is comprised of markers that can identify lung cancer among smokers as well as gradient and localized site-dependent expression patterns that recapitulate NSCLC profiles. Our findings on the adjacent field of cancerization provide additional insights into the biology of NSCLC and the development of molecular tools for the detection of the malignancy. Furthermore, we studied in detail the functional roles and properties of *LAPTM4B*, a putative oncogene that we identified as a field cancerization marker, in lung cancer pathogenesis. We found that *LAPTM4B* promotes growth of lung cancer cells in soft agar, is associated with smoking and worse overall survival in NSCLC, protects cancer cells from serum deprivation-induced growth inhibition and positively controls the autophagic response and the NRF2 transcription factor during cellular stress. Furthermore, we began to characterize the molecular spatial map of field effects that transverse the bronchus adjacent to tumors up to the nasal epithelium. This novel analysis identified field of injury/cancerization pathways and gene sets, in patients with lung cancer, which decrease in enrichment with larger distance from the tumor as well as those that persist up to the nasal epithelium. These data point to specific tumor-associated profiles that are enriched in the nasal epithelium and, thus, comprise readily accessible markers for lung cancer detection that will be further refined and validated in the next year of the funding period.

We have leveraged information about transcriptional regulation to predict that the dysregulation of the MYC and TP53 pathways are early events in SCC carcinogenesis, occurring in both premalignant lesions and tumor cells. Furthermore, we have experimentally validated that the localization of MYC shifts primarily from the cytoplasm in normal basal cells of the airway to the nucleus of premalignant lesions and tumor cells, implicating the MYC pathway as a possible target for lung cancer chemoprevention.

References

1. Siegel R, Naishadham D, Jemal A. Cancer statistics, 2013. CA Cancer J Clin 2013;63(1):11-30.
2. Services USDoHaH. The Health Consequences of Smoking. A Report of the U.S. Surgeon General. 2004.
3. National Lung Screening Trial Research T, Aberle DR, Adams AM*, et al.* Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 2011;365(5):395-409.
4. Aberle DR, DeMello S, Berg CD*, et al.* Results of the two incidence screenings in the National Lung Screening Trial. N Engl J Med 2013;369(10):920-31.
5. Spira A, Beane J, Shah V*, et al.* Effects of cigarette smoke on the human airway epithelial cell transcriptome. Proc Natl Acad Sci U S A 2004;101(27):10143-8.
6. Spira A, Beane JE, Shah V*, et al.* Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. Nat Med 2007;13(3):361-6.
7. Sridhar S, Schembri F, Zeskind J*, et al.* Smoking-induced gene expression changes in the bronchial airway are reflected in nasal and buccal epithelium. BMC Genomics 2008;9:259.
8. Beane J, Sebastiani P, Whitfield TH*, et al.* A prediction model for lung cancer diagnosis that integrates genomic and clinical features. Cancer Prev Res (Phila Pa) 2008;1(1):56-64.
9. Bhutani M, Pathak AK, Fan YH*, et al.* Oral epithelium as a surrogate tissue for assessing smoking-induced molecular alterations in the lungs. Cancer Prev Res (Phila Pa) 2008;1(1):39-44.
10. Li Y, Iglehart JD, Richardson AL, Wang ZC. The amplified cancer gene LAPTM4B promotes tumor growth and tolerance to stress through the induction of autophagy. Autophagy 2012;8(2):273-4.
11. Li Y, Zhang Q, Tian R*, et al.* Lysosomal transmembrane protein LAPTM4B promotes autophagy and tolerance to metabolic stress in cancer cells. Cancer Res 2011;71(24):7481-9.
12. Li Y, Zou L, Li Q*, et al.* Amplification of LAPTM4B and YWHAZ contributes to chemotherapy resistance and recurrence of breast cancer. Nat Med 2010;16(2):214-8.
13. Irizarry RA, Hobbs B, Collin F*, et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 2003;4(2):249-64.
14. Sporn MB, Liby KT. NRF2 and cancer: the good, the bad and the importance of context. Nat Rev Cancer 2012;12(8):564-71.
15. Kweon MH, Adhami VM, Lee JS, Mukhtar H. Constitutive overexpression of Nrf2-dependent heme oxygenase-1 in A549 cells contributes to resistance to apoptosis induced by epigallocatechin 3-gallate. J Biol Chem 2006;281(44):33761-72.
16. Brunelli M, Bria E, Nottegar A*, et al.* True 3q chromosomal amplification in squamous cell lung carcinoma by FISH and aCGH molecular analysis: impact on targeted drugs. PloS one 2012;7(12):e49689.
17. Partridge M, Kiguwa S, Langdon JD. Frequent deletion of chromosome 3p in oral squamous cell carcinoma. European journal of cancer Part B, Oral oncology 1994;30B(4):248-51.
18. Bass AJ, Watanabe H, Mermel CH*, et al.* SOX2 is an amplified lineage-survival oncogene in lung and esophageal squamous cell carcinomas. Nature genetics 2009;41(11):1238-42.
19. Bach PB, Kattan MW, Thornquist MD*, et al.* Variations in lung cancer risk among smokers. J Natl Cancer Inst 2003;95(6):470-8.
20. Byers LA, Wang J, Nilsson MB*, et al.* Proteomic profiling identifies dysregulated pathways in small cell lung cancer and novel therapeutic targets including PARP1. Cancer Discov 2012;2(9):798-811.

Appendices

Profiling premalignant lesions in lung squamous cell carcinomas identifies mechanisms involved in stepwise carcinogenesis

Aik T. Ooi[1], Adam C. Gower[11], Kelvin X. Zhang[4], Jessica L. Vick[11], Longsheng Hong[3], Brian Nagao[3], William D. Wallace[3], David A. Elashoff[5,8], Tonya C. Walser[6,7], Steven M. Dubinett[3,6,7], Matteo Pellegrini[9,10], Marc E. Lenburg[11], Avrum Spira[11], Brigitte N. Gomperts[1,2,7,9]

[1]Mattel Children's Hospital, Department of Pediatrics, Department of [2]Pulmonary Medicine, [3]Pathology and Laboratory Medicine, [4]Biological Chemistry, Howard Hughes Medical Institute, [5]Biostatistics, [10]Molecular Cell and Developmental Biology, [6]Division of Pulmonary and Critical Care Medicine, David Geffen School of Medicine, [7]The Lung Cancer Research Program of the [8]Jonsson Comprehensive Cancer Center, [9]Broad Stem Cell Research Center, University of California, Los Angeles, California; [11]Section of Computational Biomedicine, Department of Medicine, Boston University School of Medicine, Boston, Massachusetts, USA.

Running title: Identifying pathways in lung carcinogenesis

Keywords: premalignant lesions, lung cancer, stepwise carcinogenesis, RNA-seq, cancer transcriptome

Funding sources: This work was supported by Department of Defense (DOD) CTRA LC090615 to SMD, AS, BNG and TCW, Tobacco-Related Disease Research Program (TRDRP) (19FT-0046) to ATO, National Heart. Lung and Blood Institute (NHLBI) R01HL094561-01 to BNG, CIRM RN2-009-04 to BNG, the National Cancer Institute (NCI) (#U01-CA152751) to SMD,

AS, BNG, TCW, Department of Defense (DOD) (#W81XWH-10-1-1006) to SMD, AS and TCW, Department of Veteran Affairs (VA) (#5I01BX000359) to SMD and TCW, and TRDRP (#20KT-0055) to TCW.

Corresponding author: Brigitte N. Gomperts, 10833 Le Conte Ave. A2-410MDCC, Los Angeles CA 90095. 310-206-0772. bgomperts@mednet.ucla.edu

Disclosure statement: AS is a founder and consultant to Allegro Diagnostics, Inc.

Word count: 3550

Total number of figures and tables: 4

Abstract

Lung squamous cell carcinoma (SCC) is thought to arise from premalignant lesions in the airway epithelium, therefore studying these lesions is critical for understanding lung carcinogenesis. We performed RNA sequencing on laser-microdissected representative cell populations along the SCC pathological continuum of patient-matched normal basal cells, premalignant lesions, and tumor cells. We discovered transcriptomic changes and identified genomic pathways altered with initiation and progression of SCC within individual patients. We used immunofluorescent staining to confirm gene expression changes in premalignant lesions and tumor cells, including increased expression of SLC2A1, CEACAM5, and PTBP3 at the protein level and increased activation of MYC via nuclear translocation. Cytoband enrichment analysis revealed coordinated loss and gain of expression in chromosome 3p and 3q regions, respectively, during carcinogenesis. This is the first gene expression profiling of airway premalignant lesions with patient-matched samples that provides insight into the mechanisms of stepwise lung carcinogenesis.

Significance

Previous microarray and sequencing studies designed to discover early biomarkers and therapeutic targets for lung SCC had limited success identifying key driver events in lung carcinogenesis, mostly due to the cellular heterogeneity of patient samples examined and the inter-individual variability associated with difficult to obtain airway premalignant lesions and appropriate normal control samples within the same patient. Our study provides much needed information about the biology of premalignant lesions and the molecular changes that occur during stepwise carcinogenesis of SCC, and it highlights a novel approach for identifying some

of the earliest molecular changes associated with initiation and progression of lung carcinogenesis within individual patients.

Introduction

Lung cancer is the most deadly cancer worldwide, accounting for more deaths than prostate cancer, breast cancer, pancreatic cancer, and colon cancer combined (1). SCC is a common type of non-small cell lung cancer (NSCLC) that accounts for 30% of all lung cancers and is frequently associated with smoking (2). In general, despite current therapeutic strategies of chemotherapy, radiation therapy, and trials with targeted therapies, the overall survival of patients with lung cancer, including SCC, is still very poor with a five-year survival rate of 15.9% (3).

SCC often arises centrally from a large airway, usually a bronchus. Ongoing injury of airway epithelia leads to repair and regeneration that can give rise to a phenotype of squamous metaplasia and subsequently to dysplasia, both of which are histologic features seen in the airways of smokers (4, 5). It is believed that SCC develops through a series of genetic and epigenetic changes that alter the epithelium from squamous metaplasia, then to dysplasia, carcinoma *in situ* and finally to invasive carcinoma (6).

Although there have been studies devoted to discovering the genetic and molecular changes observed in lung cancer, few studies have directly investigated changes associated with squamous metaplasia or dysplasia (7-9). In fact, it is not known with certainty whether premalignant lesions of the airway are the direct progenitors of invasive SCC. To better

understand the process of carcinogenesis leading to SCC, especially those steps involved in the early and precancerous stages, a precise study of the biology of premalignant lesions is needed. Basal cells (BC) of the airway are known to be stem/progenitor cells required for airway epithelial repair (10), and we hypothesize that premalignant lesions arise from aberrant repair in these cells (11). Therefore, we profiled airway BC, premalignant lesions and tumors from the same patients to improve our understanding of the stepwise carcinogenesis in SCC and to aid in the identification of new diagnostic and therapeutic approaches for SCC and novel chemopreventive strategies.

Results

Study population, sample acquisition, and sequence alignment

Fresh frozen tissue blocks were obtained from four individuals with lung SCC (patients 1-4) at the time of tumor resection, and regions of normal BC, premalignant (squamous metaplastic and dysplastic) cells, and tumor cells were successfully captured from sectioned tissues by laser microdissection (Supplementary Fig. S1). The demographic information and clinical characteristics of these patients, as well as a description of the histology of each microdissected premalignant region, are presented in Supplementary Table S1. Sequencing libraries of the expected concentration and cDNA size ranges were generated from RNA isolated from the microdissected cells. All sequenced samples produced reads with mean Phred quality scores above 25, indicating that it was possible to generate sequencing libraries of good quality with our method of isolating RNA from laser-microdissected materials.

A table of the number of reads that aligned uniquely within each sample is shown in Supplementary Table S2. The fraction of reads aligning to the mitochondrial genome varied considerably among samples. In patients 1 and 2, this fraction varied from 7% to 28%, but in patients 3 and 4, mitochondrial reads comprised from 22 to 65% of uniquely aligned reads, with the highest fraction found in the tumor samples from patient 4 (57-65%). Because of this large amount of variability, reads aligning to the mitochondrial genome were discarded from analysis after alignment, and RPKM (reads per kilobase per millions of reads) values were computed relative to the total number of reads aligning uniquely to the nuclear genome.

Identification of genes associated with carcinogenesis

To identify SCC-associated genes whose expression is also associated with progression from normal airway BC to premalignant (metaplastic or dysplastic) lesions, a multi-step procedure was used as outlined in Fig. 1A. First, Ensembl Genes with zero aligned reads in all samples from at least one patient were removed from analysis (to ensure that all patients contributed evidence to each result), leaving 20,817 genes for analysis. This list was then filtered to consider only those genes with substantial evidence of expression (median of greater than 50 uniquely aligned reads across all samples), leaving 7,025 genes for analysis. Using linear mixed-effects models and negative binomial generalized linear models (see Supplementary Methods for details), we then identified 626 early-stage genes (significantly differentially expressed in a similar manner in both premalignant lesions and tumor compared to normal BC), 730 late-stage genes (significantly differentially expressed in a similar manner in tumor compared to both premalignant lesions and normal BC), and 68 "stepwise" genes (significantly differentially expressed in both of the described stages of carcinogenesis) (Fig. 1B, Supplementary Table S3).

Experimental and computational validation of candidate genes

Three genes were selected for further validation: *CEACAM5, SLC2A1* and *PTBP3*. These genes, whose expression was upregulated in premalignant lesions and tumor cells compared to normal BC, were chosen because of their potential roles in the biology of lung carcinogenesis. The expression of *CEACAM5* and *SLC2A1* was measured by performing qPCR on remaining material from the sequencing libraries of patients 3 and 4, as well as on laser-microdissected RNA from four additional independent cases (patients 5-8). In each case, the mRNA level of each gene was significantly higher (sign test $p < 0.05$) in the premalignant lesion than in normal BC (Fig. 2A). Because mRNA and protein levels may not always be well correlated (12-14), immunofluorescent staining was performed in sections of normal epithelium, premalignant lesion, and carcinoma from two independent cases (patients 9 & 10). CEACAM5 and SLC2A1 were not detectable in the normal epithelia, but they were highly expressed in cells within both metaplastic lesions and the SCC tumors (Fig. 2B & 2C). SLC2A1 was expressed throughout the KRT5+ component of the tumor, whereas CEACAM5 was expressed in some, but not all, KRT5+ tumor cells. PTBP3 was strongly expressed in premalignant lesions and tumor cells, and although it was strongly expressed in columnar KRT5- cells of normal airway epithelium, its expression was undetectable in normal BC (Supplementary Fig. S2).

To better understand the biological role these genes may play in the development of lung SCC, the Gene Expression Omnibus (GEO) Profiles tool was used to examine their expression in GEO DataSets associated with experimental parameters relevant to lung SCC carcinogenesis. First, *SLC2A1* and *PTBP3* were confirmed to be significantly upregulated (*SLC2A1*: $p = 0.004$;

PTBP3: $p = 0.017$) in an independent set of SCC tumors with respect to paired samples of adjacent normal tissue (GEO DataSet GDS1312) (15); however, the expression of *CEACAM5* was unchanged ($p = 0.64$). Next, a collection of SCC and adenocarcinoma (ADC) lung tumors (GDS3627) (16) was interrogated to determine the specificity of the expression of these genes with respect to the SCC tumor type. The expression of *SLC2A1* and *PTBP3* were again strongly increased in SCC tumors compared with ADC tumors (*SLC2A1*: $p = 1.1 \times 10^{-7}$; *PTBP3*: $p = 0.0004$); however, *CEACAM5* was moderately downregulated in SCC relative to ADC ($p = 0.08$). Finally, because premalignant lesions in large central airways are believed to arise from injury caused by cigarette smoking, the expression levels of these genes were examined in a study of bronchoscopic brushings of healthy current, former, and never smokers (GDS534) (17). In this study, *CEACAM5* and *SLC2A1* were significantly upregulated in brushings from current smokers compared with those from never smokers (*CEACAM5*: $p = 0.0001$; *SLC2A1*: $p = 0.016$), although *PTBP3* was not ($p = 0.66$).

Prediction of chromosomal gains and losses during carcinogenesis

Gene Set Enrichment Analysis (GSEA) performed using positionally defined gene sets (cytobands) revealed that late-stage (but not early-stage) carcinogenesis is associated with a coordinate loss of expression in the p arm of chromosome 3 and an attendant gain of expression in 3q26.33-3q29 (Fig. 3A), which corresponds to previously reported observations of frequent 3p deletion and 3q amplification in squamous tumors (18, 19). In particular, chromosomal band 3q26.33 has been reported to be consistently amplified in lung SCC (20).

Identification of biological changes in early- and late-stage carcinogenesis

Ingenuity Pathway Analysis (IPA) (Ingenuity Systems) was used to further characterize the changes in biological functions resulting from the differential expression of genes associated with early-stage events, which contribute to the initiation and formation of premalignant lesions, or with late-stage events, which are involved in the progression from premalignant lesions to tumor. This analysis revealed that the early-stage carcinogenesis was characterized uniquely by increased protein ubiquitination and cell cycle progression, whereas the late-stage events were marked primarily by increased transcriptional and translational activity and cellular migration and transformation (Fig. 3B, Supplementary Table S4). In addition, an increase in cell survival and proliferation and a corresponding downregulation of cell death mechanisms was observed throughout both stages of carcinogenesis.

IPA was also used to determine whether the genes identified to be differentially expressed either early or late in carcinogenesis are enriched in known targets of various transcription factors. This approach revealed that the set of genes that is differentially expressed early in carcinogenesis and remains dysregulated in tumor cells is enriched in previously reported targets of MYC and TP53 (Fig. 4A, Supplementary Table S5 & S6). As MYC and TP53 are predicted to activate or repress the expression of these targets, respectively, this suggests that MYC activity is significantly induced ($p = 2.41\text{x}10^{-5}$, z-score = 3.789) and TP53 activity is potentially repressed ($p = 9.30\text{x}10^{-8}$, z-score = -1.034) during early carcinogenesis, and that their activity remains altered throughout tumorigenesis. Importantly, the gene expression levels of *TP53* and *MYC* did not change significantly with respect to the pathological continuum from normal to tumor, suggesting that the predicted changes in their activity are due to post-transcriptional regulation.

To test the hypothesis that MYC activity is induced during early SCC carcinogenesis, immunofluorescent staining of MYC was performed to examine its nuclear and cytoplasmic localization in normal BC, premalignant lesions, and tumor cells (Fig. 4B). MYC staining was exclusive to the nuclei of premalignant lesions and tumor cells. In the histologically normal BC of the airways from patients with lung cancer, however, MYC was localized predominantly in the cytoplasm, although some areas of nuclear staining were also seen. The increased expression of MYC targets in the premalignant lesions and tumor cells, together with a concomitant increase in the nuclear localization of MYC, is strong evidence for a carcinogenesis-associated increase in MYC activity without a significant increase in gene expression.

Discussion

Little is known about the development of premalignant lesions and their progression to SCC because of a lack of appropriate *in vitro* and *in vivo* stepwise models of SCC tumorigenesis. The current practice of profiling whole-tissue biopsies is not conducive to studying premalignancy, as such biopsy samples are highly heterogeneous (7-9) and are therefore subject to confounding cell type-specific effects. The approach described here allows the examination of specific cell populations along the continuum of lung carcinogenesis and the study of relationships between each of these populations. Furthermore, as the premalignant lesions are in close proximity to SCC within the same patients, it is reasonable to expect that alterations in gene expression shared between premalignant and tumor cells reflect molecular changes that occur during carcinogenesis.

We focused specifically on the expression patterns of three genes, *CEACAM5*, *SLC2A1*, and *PTBP3*, that are upregulated in the premalignant lesions (and, in the case of *SLC2A1*, further upregulated in tumor cells). CEACAM5, a cell surface glycoprotein that plays a role in cell adhesion and intracellular signaling, has been shown to be important in other epithelial cell cancers, such as colon cancer (21). SLC2A1 (also known as glucose transporter 1, or GLUT1) is a facilitative glucose transporter associated with hepatocellular cancer and head and neck squamous cell carcinoma (22, 23). PTBP3 (also known as regulator of differentiation 1, or ROD1) is an RNA-binding protein that regulates pre-mRNA alternative splicing and plays a role in the regulation of cell proliferation and differentiation (24, 25). The protein level expression of each gene was substantially increased in premalignant lesions and tumor cells, although the expression of CEACAM5 within the tumor cells was more heterogeneous than that of the other genes. Additionally, although PTBP3 was strongly expressed in normal airway epithelium, its expression was restricted to columnar KRT5- cells. It is worth noting that, as the columnar epithelial cells are in close proximity to the BC, it would be almost impossible to identify the increased expression of *PTBP3* in premalignant lesions and tumors without laser microdissection of the normal BC from the airway epithelium.

We also examined the expression of these genes in publicly available microarray datasets related to SCC carcinogenesis. In one such experiment, the genes *SLC2A1* and *PTBP3* were significantly upregulated in lung SCC tumors relative to matched adjacent normal tissue, but unexpectedly, *CEACAM5* was not. However, that study profiled tumor biopsies, which often contain significant stromal contamination; moreover, we observed substantial heterogeneity of CEACAM5 immunostaining in SCC tumor tissue in this study. The identification of *CEACAM5*

as an early-stage marker of squamous lung carcinogenesis in this study may therefore be attributable to the careful laser microdissection of SCC tumor cells from the surrounding stroma. Finally, because lung SCC is strongly associated with a history of tobacco smoking, we examined the relationship between smoking history and the expression of these genes in a previous study of bronchoscopic brushings. In that study, *CEACAM5* and *SLC2A1* were significantly upregulated in brushings from current smokers compared with those from never smokers. In a subsequent study from the same authors (26), the expression of *CEACAM5* was reported to be irreversibly altered in former smokers for up to several decades after smoking cessation, suggesting that a smoking-associated increase in *CEACAM5* expression in histologically normal airway epithelium may be an early event associated with carcinogenesis in these individuals.

We used GSEA to identify chromosomal regions that were enriched in differentially expressed genes, which suggested that the frequent 3p loss and 3q amplification that are characteristic of SCC (and rare in ADC) (20, 27) are late-stage events in SCC carcinogenesis. We also used IPA to identify biological functions and regulators that were overrepresented among the genes associated with early- or late-stage carcinogenesis. This analysis revealed that early-stage carcinogenesis is marked primarily by increased flux through the cell cycle, but that cellular proliferation continues throughout late-stage carcinogenesis. Finally, we used IPA to make predictions about the upstream regulators that might be responsible for these changes, and identified the transcription factors TP53 and MYC as likely candidates based on the coordinate differential expression of their target genes. Accordingly, immunofluorescent staining revealed that MYC is found in the cytoplasm in most normal BC, but is localized in the nuclei of

premalignant lesions and tumor cells, suggesting that activation of MYC by nuclear translocation could be an important event contributing to dysregulated cell cycle progression during SCC carcinogenesis. Previous reports have also identified the potential importance of MYC in premalignant lesions of lung carcinoma (28) and breast cancer (29).

While this study represents a novel approach for identifying driver molecular events associated with squamous cell lung carcinogenesis, there are a number of important limitations to the work. Our model assumes that there is a molecular relationship between the premalignant and tumor cells found within the same patient's airway, although the lesions may develop from disparate clonal populations, thereby limiting the interpretation of those changes as reflecting a stepwise change between lesions. Longitudinal studies of premalignant lesions resampled over time are needed to identify molecular alterations associated with progression or regression within a clonal population of cells. Furthermore, our group and others have previously reported molecular alterations throughout the histologically normal airway of smokers with lung cancer (30). Those molecular events in the histologically normal "field of injury" may reflect some of the earliest events in carcinogenesis and will not be captured directly by our approach.

In summary, we present a novel and technically challenging method to study the biology of premalignant lesions and carcinogenesis in lung SCC. Our analysis identified coordinate changes in the activity of upstream regulators and the expression of downstream genes within the same patient during early- and late-stage carcinogenesis. Further work will be necessary to determine if any of these genes can also be used to distinguish premalignant lesions that will progress to

cancer from those that will regress. Genes identified and validated in this manner might serve as early biomarkers for SCC detection and targets for SCC chemoprevention.

Materials and Methods

Case selection and histology review

SCC cases were reviewed with pathologists to identify regions of normal airway epithelium, squamous metaplasia or dysplasia, or carcinomas. Patients with fresh frozen or formalin-fixed paraffin-embedded (FFPE) tissue blocks containing all three regions were selected for the study. Immunofluorescent staining of KRT5 was performed to validate the identification of selected lesions. Fresh frozen tissues were used for RNA sequencing, whereas FFPE tissues were used for validation of independent cases with quantitative real-time PCR (qPCR) and immunofluorescent staining.

Laser Capture Microdissection (LCM)

Tissues were sectioned at a 7-micron thickness and mounted on regular uncharged glass slides for patients 1, 2, and 3, and on polyethylene napthalate (PEN) membrane slides (Leica) for patient 4, followed by H&E staining. LCM was performed using the Arcturus eIIx for patient 1 and 2, Zeiss PALM for patient 3, and Leica LMD7000 for patient 4. A tissue area of 800,000 to 1,200,000 μm^2 was dissected and collected from each lesion.

RNA extraction and sequencing library preparation

RNA was extracted from laser-microdissected cells using the RNeasy Micro Kit (QIAGEN). The cDNA was generated using the Ovation RNA-Seq System (NuGEN) for patients 1 and 2 and the

Ovation RNA-Seq V2 System (NuGEN) for patients 3 and 4. For patients 1 and 2, cDNA of ~200 bp was selected by gel purification. For patients 3 and 4, the cDNA was sheared to 140–180 bp using the Covaris focused-ultrasonicator with the following settings: duty cycle 10%; intensity 5; cycles per burst 200; total time 6 minutes. The size range of the sheared cDNA was confirmed by Bioanalyzer analysis prior to library construction using the Encore Library System (NuGEN). The average size of each library was estimated by Bioanalyzer analysis, and the concentration of each was measured on the Qubit fluorometer (Invitrogen).

Sequence analysis

Sequencing libraries from patients 1 and 2 were each sequenced on a single flow cell lane of an Illumina Genome Analyzer IIx, generating 36-base single-end reads, and libraries from patients 3 and 4 were each sequenced on a single flow cell lane of an Illumina HiSeq 2000, generating 50-base single-end reads. All reads were trimmed to 35 bases before alignment. In the case of patient 1, the first base of each read was also trimmed off due to a problem with the first sequencing cycle. Reads that failed Illumina's chastity filter [brightest intensity / (brightest intensity + second brightest intensity) < 0.6 for at least two of the first 25 cycles] were automatically removed during preprocessing. The remaining reads were aligned to the human genome (build hg19) using Bowtie v0.12.7 (31), allowing only unique alignments and up to two mismatches per read. Reads aligning to the mitochondrial genome were removed from further analysis. Gene expression estimates were then computed by measuring the coverage of each of 55,841 Ensembl Gene loci (Ensembl build 69) using the BEDTools software suite (32). The coverage for each Ensembl Gene locus in each sample was then normalized to the size of the locus and the total number of reads mapping uniquely to the nuclear genome to obtain an RPKM

(33) value for each gene in each sample. RPKM values were seventh-root-transformed prior to analysis to produce an approximately normal distribution of (nonzero) gene expression values.

Statistical analysis

All models were created using the R environment for statistical computing (version 2.12.0) (34). Linear mixed-effects models were created using the *nlme* R package (version 3.1-97) (35) and negative binomial models were created using the *MASS* R package (36).

Gene Set Enrichment Analysis (GSEA)

Positionally defined (cytoband) Ensembl Gene sets were created using the *biomaRt* R package (37) to extract chromosomal band annotation for Ensembl Gene identifiers using Ensembl version 69. These gene sets were then used to perform pre-ranked GSEA (38), using lists of all Ensembl Genes ranked by the *t* statistics from the linear mixed-effects models, to identify cytobands that were overrepresented among genes coordinately up- or down-regulated in premalignant or tumor cells compared with normal BC. Analysis was performed using GSEA v2.0.8 (build 14) with 1000 permutations, removal of gene sets with > 500 genes, and a random seed of 1234.

Quantitative Real-Time Polymerase Chain Reaction (qPCR)

Amplified cDNA generated during the library preparation for patients 3 and 4 was used for qPCR analysis. In addition, normal BC and premalignant lesions from four independent patients were laser-microdissected from FFPE tissues, and cDNA was generated using the Ovation RNA-Seq FFPE System. TaqMan Gene Expression Assays (Life Technologies) were used to examine

the expression levels of selected candidate genes (*CEACAM5*, *SLC2A1*) in normal BC and premalignant lesions. β2-microglobulin (*B2M*) was used as an endogenous control. Statistical analysis was performed using the sign test.

Immunofluorescent staining

FFPE tissues were sectioned at a 5-micron thickness and stained as previously described (11). Antibodies used are: rabbit anti-KRT5, mouse anti-KRT5, mouse anti-PTBP3, rabbit anti-c-myc (Abcam); mouse anti-CEACAM5 (ProMab Biotechnologies Inc); rabbit anti-SLC2A1 (Alpha Diagnostic International); anti-rabbit Cy3, anti-mouse Alexafluor 647, anti-mouse Cy3, anti-rabbit Alexafluor 647 (Jackson ImmunoResearch). Immunostained tissues were visualized on an Axiocam system (Zeiss), and images were taken using the Axiovision software.

Acknowledgments

We would like to thank Gang Liu and Lingqi Luo for performing the RNA-seq of patients 1, 2 and 4, UCLA Broad Stem Cell Research Center High-Throughput Sequencing Core Resource for performing the RNA-seq for patient 3, as well as Frank Schembri for technical assistance in optimizing RNA extraction and Josh Campbell and Nacho Caballero for assistance with negative binomial analysis. Histological services were performed by the UCLA Translational Pathology Core Laboratory (TPCL) and laser microdissection was performed through the TPCL and the Advanced Light Microscopy Core. The statistical analysis for qPCR was performed by Tristan Grogan and supported by NIH/National Center for Advancing Translational Science (NCATS) UCLA CTSI Grant Number UL1TR000124.

Figure legends

Figure 1. Identification of genes associated with early- or late-stage SCC carcinogenesis. **A**. Analysis flowchart. Uniquely aligned reads were assigned to 55,841 Ensembl Gene loci (Ensembl build 69). Two statistical models were then applied to identify genes with significant ($p < 0.05$) differential expression between tumor and normal cells, as well as between premalignant and normal cells ("early" genes), between tumor and premalignant cells ("late" genes), or both ("stepwise" genes). **B**. Expression heatmap. Root-transformed RPKM values were scaled to a mean of zero and standard deviation of one within each patient; red and blue indicate genes with expression that is higher or lower than the mean within each patient, respectively. Genes are hierarchically clustered within each group (early, stepwise, late).

Figure 2. Experimental validation of *CEACAM5* and *SLC2A1* expression. **A**. Quantitative real-time PCR. Box plots represent RNA levels of *CEACAM5* and *SLC2A1* in normal BC and premalignant lesions from six patients, showing increased expression of both genes in premalignant lesions compared to normal BC. *B2M* was used as the endogenous control. **B-C**. Immunofluorescent staining of CEACAM5 and SLC2A1. Protein staining shows increased expression of CEACAM5 and SLC2A1 in premalignant lesions and SCC compared to BC in the normal epithelium. Top rows: normal airway epithelium; middle rows: premalignant lesions; bottom rows: SCC. Left columns: KRT5, stained in red as marker for BC, premalignant lesions, and tumor cells; middle columns: SLC2A1 or CEACAM5, stained in green; right columns: merged images of left and middle columns. DAPI, stained in blue, as nuclear marker. White scale bars: 50 μm. Insets show close-up views of the boxed regions.

Figure 3. Identification of coordinately regulated chromosomal regions and pathways. **A**. Identification of differentially regulated cytobands by GSEA. Positional sets of Ensembl Genes (cytobands) were obtained from Biomart and used to perform pre-ranked GSEA with lists of *t* statistics (P vs N, T vs P) from the linear mixed-effects models. Dashed lines indicate nominal $p = 0.05$. **B.** Identification of dysregulated biological functions by IPA. Selected biological functions ($p < 0.05$ and z-scores ≥ 2 or ≤ -2) predicted to be significantly increased (positive x-axis values) or decreased (negative x-axis values) in early-stage (magenta bars) and late-stage (yellow bars) carcinogenesis.

Figure 4. Identification of MYC as a dysregulated transcription factor in early carcinogenesis. **A**. Identification of dysregulated transcription factors by IPA. Expression patterns of known downstream targets of TP53 and MYC suggest the inhibition of TP53 and the activation of MYC during early carcinogenesis. **B**. Experimental validation of MYC activation in premalignant lesions and SCC. Immunofluorescent staining shows detection of MYC expression in the nuclei of premalignant lesions and SCC compared to cytoplasmic localization in BC. Top rows: normal airway epithelium; middle rows: premalignant lesion; bottom rows: SCC. Left columns: KRT5, stained in green as marker for BC, premalignant lesion, and tumor cells; middle columns: MYC, stained in red; right columns: merged images of left and middle columns. DAPI, stained in blue, as nuclear marker in all images. White scale bars: 50 μm. Insets show close-up views of the boxed regions.

References

1. Siegel R, Naishadham D, Jemal A. Cancer statistics, 2012. CA Cancer J Clin.62:10-29.

2. Stat bite: Mortality from lung and bronchus cancer by race/ethnicity, 1998-2002. J Natl Cancer Inst. 2006;98:158.

3. Ettinger DS, Akerley W, Borghaei H, Chang AC, Cheney RT, Chirieac LR, et al. Non-small cell lung cancer. Journal of the National Comprehensive Cancer Network : JNCCN. 2012;10:1236-71.

4. Colby TV, Wistuba, II, Gazdar A. Precursors to pulmonary neoplasia. Advances in anatomic pathology. 1998;5:205-15.

5. Kerr KM. Pulmonary preinvasive neoplasia. Journal of clinical pathology. 2001;54:257-71.

6. Peebles KA, Lee JM, Mao JT, Hazra S, Reckamp KL, Krysan K, et al. Inflammation and lung carcinogenesis: applying findings in prevention and treatment. Expert Rev Anticancer Ther. 2007;7:1405-21.

7. Kettunen E, Anttila S, Seppanen JK, Karjalainen A, Edgren H, Lindstrom I, et al. Differentially expressed genes in nonsmall cell lung cancer: expression profiling of cancer-related genes in squamous cell lung cancer. Cancer Genet Cytogenet. 2004;149:98-106.

8. Seo JS, Ju YS, Lee WC, Shin JY, Lee JK, Bleazard T, et al. The transcriptional landscape and mutational profile of lung adenocarcinoma. Genome Res. 2012.

9. Xi L, Feber A, Gupta V, Wu M, Bergemann AD, Landreneau RJ, et al. Whole genome exon arrays identify differential expression of alternatively spliced, cancer-related genes in lung cancer. Nucleic Acids Res. 2008;36:6535-47.

10. Hong KU, Reynolds SD, Watkins S, Fuchs E, Stripp BR. Basal cells are a multipotent progenitor capable of renewing the bronchial epithelium. Am J Pathol. 2004;164:577-88.

11. Ooi AT, Mah V, Nickerson DW, Gilbert JL, Ha VL, Hegab AE, et al. Presence of a putative tumor-initiating progenitor cell population predicts poor prognosis in smokers with non-small cell lung cancer. Cancer Res. 2010;70:6639-48.

12. Greenbaum D, Colangelo C, Williams K, Gerstein M. Comparing protein abundance and mRNA expression levels on a genomic scale. Genome Biol. 2003;4:117.

13. Taniguchi Y, Choi PJ, Li GW, Chen H, Babu M, Hearn J, et al. Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. Science. 2010;329:533-8.

14. Gygi SP, Rochon Y, Franza BR, Aebersold R. Correlation between protein and mRNA abundance in yeast. Molecular and cellular biology. 1999;19:1720-30.

15. Wachi S, Yoneda K, Wu R. Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. Bioinformatics. 2005;21:4205-8.

16. Kuner R, Muley T, Meister M, Ruschhaupt M, Buness A, Xu EC, et al. Global gene expression analysis reveals specific patterns of cell junctions in non-small cell lung cancer subtypes. Lung Cancer. 2009;63:32-8.

17. Spira A, Beane J, Shah V, Liu G, Schembri F, Yang X, et al. Effects of cigarette smoke on the human airway epithelial cell transcriptome. Proc Natl Acad Sci U S A. 2004;101:10143-8.

18. Brunelli M, Bria E, Nottegar A, Cingarlini S, Simionato F, Calio A, et al. True 3q chromosomal amplification in squamous cell lung carcinoma by FISH and aCGH molecular analysis: impact on targeted drugs. PLoS One. 2012;7:e49689.

19. Partridge M, Kiguwa S, Langdon JD. Frequent deletion of chromosome 3p in oral squamous cell carcinoma. European journal of cancer Part B, Oral oncology. 1994;30B:248-51.

20. Bass AJ, Watanabe H, Mermel CH, Yu S, Perner S, Verhaak RG, et al. SOX2 is an amplified lineage-survival oncogene in lung and esophageal squamous cell carcinomas. Nat Genet. 2009;41:1238-42.

21. Pignatelli M, Durbin H, Bodmer WF. Carcinoembryonic antigen functions as an accessory adhesion molecule mediating colon epithelial cell-collagen interactions. Proc Natl Acad Sci U S A. 1990;87:1541-5.

22. Amann T, Maegdefrau U, Hartmann A, Agaimy A, Marienhagen J, Weiss TS, et al. GLUT1 expression is increased in hepatocellular carcinoma and promotes tumorigenesis. Am J Pathol. 2009;174:1544-52.

23. Heikkinen PT, Nummela M, Jokilehto T, Grenman R, Kahari VM, Jaakkola PM. Hypoxic conversion of SMAD7 function from an inhibitor into a promoter of cell invasion. Cancer Res. 2010;70:5984-93.

24. Sadvakassova G, Dobocan MC, Difalco MR, Congote LF. Regulator of differentiation 1 (ROD1) binds to the amphipathic C-terminal peptide of thrombospondin-4 and is involved in its mitogenic activity. Journal of cellular physiology. 2009;220:672-9.

25. Yamamoto H, Tsukahara K, Kanaoka Y, Jinno S, Okayama H. Isolation of a mammalian homologue of a fission yeast differentiation regulator. Molecular and cellular biology. 1999;19:3829-41.

26. Beane J, Sebastiani P, Liu G, Brody JS, Lenburg ME, Spira A. Reversible and permanent effects of tobacco smoke exposure on airway epithelial gene expression. Genome Biol. 2007;8:R201.

27. Massion PP, Kuo WL, Stokoe D, Olshen AB, Treseler PA, Chin K, et al. Genomic copy number analysis of non-small cell lung cancer using array comparative genomic hybridization: implications of the phosphatidylinositol 3-kinase pathway. Cancer Res. 2002;62:3636-40.

28. Massion PP, Zou Y, Uner H, Kiatsimkul P, Wolf HJ, Baron AE, et al. Recurrent genomic gains in preinvasive lesions as a biomarker of risk for lung cancer. PLoS One. 2009;4:e5611.

29. Ling H, Sylvestre JR, Jolicoeur P. Notch1-induced mammary tumor development is cyclin D1-dependent and correlates with expansion of pre-malignant multipotent duct-limited progenitors. Oncogene. 2010;29:4543-54.

30. Spira A, Beane JE, Shah V, Steiling K, Liu G, Schembri F, et al. Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. Nature medicine. 2007;13:361-6.

31. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10:R25.

32. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841-2.

33. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature methods. 2008;5:621-8.

34. R Development Core Team. R: A langudepot and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria.; 2011.

35. Pinheiro J, Bates D, DebRoy S, Sarkar D, and R Development Core Team. nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-97.

36. Venables W, Ripley, BD. Modern Applied Statistics with S. Fourth Edition. Springer, New York; 2002.

37. Durinck S, Huber W. biomaRt: Interface to BioMart databases (e.g. Ensembl, COSMIC ,Wormbase and Gramene). R package version 2.6.0.

38. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102:15545-50.

miR-4423 is a Primate-Specific Regulator of Airway Epithelial Cell Differentiation and Lung Carcinogenesis

Catalina Perdomo[1,2], Joshua D. Campbell[1,3], Joseph Gerrein[1,3], Carmen Tellez[4], Carly B. Garrison[1,2], Tonya C. Walser[7], Eduard Drizik[1], Huiqing Si[1], Adam C. Gower[1,3], Jessica Vick[1,2], Christina Anderlind[1], George R. Jackson[6], Courtney Mankus[6], Frank Schembri[5], Carl O'Hara[8], Brigitte N. Gomperts[7], Steven M. Dubinett[7], Patrick Hayden[6], Steven A. Belinsky[4], Marc E. Lenburg[∞1,2,3], Avrum Spira*[∞1,2,3,5]

[1] Division of Computational Biomedicine, Department of Medicine, Boston University Department of Medicine, 72 East Concord Street, Boston, MA 02118, USA.

[2] Genetics and Genomics Program, Boston University Department of Medicine, 715 Albany Street, Boston, MA 02118, USA.

[3] Bioinformatics Graduate Program, Boston University, 44 Cummington Street, Boston, MA 02215, USA.

[4] Lovelace Respiratory Research Institute, 2425 Ridgecrest Drive SE, Albuquerque, NM 87108, USA.

[5] Pulmonary Center, Boston University Department of Medicine, 715 Albany Street, Boston, MA 02118, USA.

[6] MatTek corporation, 200 Homer Avenue, Ashland, MA, 01721, USA.

[7] Division of Pulmonary and Critical Care Medicine, David Geffen School of Medicine at UCLA, Los Angeles, California 90095, USA.

[8] Department of Pathology and Laboratory Medicine, Boston University School of Medicine, 670 Albany Street, Boston, MA 02118, USA.

[∞]co-senior authors

*To whom correspondence should be addressed: aspira@bu.edu

Abstract

Smoking is a significant risk factor for lung cancer, the leading cause of cancer-related deaths worldwide. While microRNAs are regulators of many airway gene-expression changes induced by smoking, their role in modulating changes associated with lung cancer in these cells remains unknown. Here, we use next-generation sequencing of small RNAs in the airway to identify miR-4423 as a novel primate-specific microRNA associated with lung cancer and expressed primarily in mucociliary epithelium. The endogenous expression of miR-4423 increases as bronchial epithelial cells undergo differentiation into mucociliary epithelium *in vitro* and its overexpression during this process causes an increase in the number of ciliated cells. Furthermore, expression of miR-4423 is reduced in most lung tumors and in cytologically normal epithelium of the mainstem bronchus of smokers with lung cancer. In addition, ectopic expression of miR-4423 in a subset of lung cancer cell lines reduces their anchorage-independent growth and significantly decreases the size of the tumors formed in a mouse xenograft model. Consistent with these phenotypes, overexpression of miR-4423 induces a differentiated-like pattern of airway epithelium gene expression and reverses the expression of many genes that are altered in lung cancer. Together, our results indicate that miR-4423 is a novel regulator of airway epithelium differentiation and that the abrogation of its function contributes to lung carcinogenesis.

Significance

MicroRNAs are small non-coding RNAs that negatively regulate gene expression and have been implicated in a variety of cellular processes. Using small RNA sequencing, we identified miR-4423 as a novel primate-specific microRNA whose expression is largely restricted to airway epithelium and which functions as a regulator of airway epithelium differentiation and a repressor of lung carcinogenesis. Understanding miR-4423's role in airway development may provide insights into primate-specific aspects of airway biology and the evolution of primate-specific

tumor suppressors. Moreover, this study opens the possibility that microRNAs might be useful for the early detection of lung cancer in the proximal airway and that miR-4423 mimetics might also be used as therapeutic agents to specifically target lung cancer.

/body

MicroRNAs are a class of small, non-coding RNAs that reduce gene expression and protein translation through complementary binding to the 3' UTR of target genes. These small RNA species have emerged as key regulators of virtually all cellular processes, including cell growth, stress response, tissue specification and cell differentiation (1, 2). Many microRNAs are expressed in a tissue-specific manner and directly regulate genes that are important in specifying the developmental fate of the cells in which they are expressed. For example, miR-449 is expressed specifically in columnar multiciliated airway epithelial cells and promotes the differentiation of airway ciliated cell progenitors by repressing the Delta/Notch pathway (3, 4). In addition, expression of tissue-specific microRNAs is often lost during carcinogenesis and restoring their expression can promote the redifferentiation of cancer cells to their original tissue type, suggesting a potential new avenue for cancer therapy (5, 6). Differences in microRNA expression have been associated with cancer prognosis and recent findings suggest that microRNA expression measured in readily collected samples can be used for early cancer detection (7, 8).

Smoking is a significant risk factor for lung cancer, the most common cause of cancer-related deaths worldwide (9). Smoking induces molecular alterations throughout the respiratory tract, including the nasal, buccal, and bronchial epithelium (10, 11). We and others have previously characterized the effect of smoking on the bronchial epithelium transcriptome (12-15), and we have shown that microRNAs play a role in regulating these smoking-related changes in gene expression (16). We have also found that cytologically normal bronchial epithelial cells from the

mainstem bronchus of smokers with and without lung cancer have marked differences in gene expression that can serve as an early detection biomarker for lung cancer (17). These data have led us to hypothesize that, as in the case of the cellular response to smoking, microRNAs might modulate cancer-associated gene expression differences in airway epithelium. Moreover, specific patterns of microRNA expression in these cells might also be able to serve as a biomarker for lung cancer detection.

In this study, we have used sequencing of small RNAs to identify microRNAs expressed in the airway epithelium in the setting of lung cancer. We report the identification and characterization of miR-4423 as a primate-specific microRNA whose expression is largely restricted to the respiratory tract epithelium and plays a role in the development of the mucociliary epithelium by promoting ciliated cell differentiation. In addition, we show that miR-4423 is decreased in the cytologically normal bronchial epithelium of smokers with lung cancer and in lung tumors, and inhibits anchorage-independent growth in lung cancer cell lines and their ability to grow as tumors in a mouse xenograft model. Together, our results suggest that miR-4423 is involved in promoting airway differentiation and its inhibition is implicated in lung carcinogenesis.

Results

Small RNA airway transcriptome sequencing

Pools of small RNA (< 40 nt) from the bronchial airway epithelium of healthy never smokers, healthy current smokers, current or former smokers with lung cancer, and current or former smokers with benign lung disease (3 individuals per pool; see **SI Appendix** Table S1 for subject demographics) were sequenced using ABI SOLiD platform (Applied Biosystems). Each read was trimmed and aligned to hg19 using Bowtie (18) allowing up to 2 mismatches per read. On average, 67 million reads were obtained per sample, of which 33.2 million reads aligned to the genome, and 9.8 million reads mapped to a known microRNA precursor from release 16 of

miRBase (19) (see Table S2 for sequence alignment statistics). A total of 488 microRNAs had expression levels of one read per million reads (RPM) or greater in at least one sample.

Computational prediction of novel microRNAs

The miRDeep algorithm (20) was used to identify transcribed regions that were predicted to fold into a canonical microRNA structure. 54 microRNA hairpins were identified, five of which were not previously annotated as microRNAs (miRDeep score > 50 and RPM > 5; see Table S3 for a list of predicted novel microRNAs). A predicted microRNA precursor mapping to chr1: 85,599,489-85,599,545 (Figure 1A) was the top scoring miRDeep prediction that did not overlap a known microRNA in release 16 of miRBase, and the 9th highest scoring miRDeep prediction overall. Within the predicted precursor, miRDeep predicted two mature microRNAs, 3p and 5p (Figure 1B). Both forms are highly expressed in airway epithelium, with the 3p and 5p being the 85th (91st percentile) and 107th (89th percentile) most highly expressed mature microRNAs, respectively. Quantitative real-time PCR (qRT-PCR) was used to validate the expression of both forms of the putative microRNA in human airway epithelial cells from bronchial brushings (Figure S1). During the preparation of this manuscript, the sequence of this putative microRNA was deposited in miRBase release 17 as hsa-miR-4423, based on a small number of sequencing reads in studies that characterized the small RNA transcriptome of melanoma (2 reads) and cervical tumors (1 read) (21, 22). Given the low level of expression in these studies and the high level of expression in airway epithelium, we investigated the regulation, expression, and function of miR-4423 in the respiratory tract.

Conservation of miR-4423

A multiple species alignment revealed that the genomic locations of the primary transcript and precursor of miR-4423 are highly conserved in simians, i.e., the great apes (human, chimpanzee, gorilla, orangutan), Old World monkeys (rhesus, baboon), and New World monkeys (marmoset).

The regions corresponding to the miR-4423-5p and -3p seeds are perfectly conserved in simians, with the exception of a single base substitution in the miR-4423-5p seed region of the rhesus genome. In contrast, the regions of the primary transcript and precursor are less well conserved in two prosimians (mouse lemur and bushbaby) and are not conserved among non-primate mammals (Figure 1C). These results suggest that miR-4423 is a recently evolved microRNA.

miR-4423 is processed by Dicer and Argonaute in vitro

We did not detect expression of either form of miR-4423 or the primary transcript in the 13 lung cancer cell lines or 8 undifferentiated bronchial epithelial cell lines that we examined (Table S4; S5). To experimentally validate whether the identified sequence of miR-4423 encodes a functional mature microRNA, H1299 cells were transfected with a plasmid expressing the miR-4423 predicted precursor sequence (including ≈200 bp of flanking region) under the control of the CMV promoter, which resulted in the expression of mature miR-4423-3p and -5p (Figure S2). Expression of both mature forms was significantly reduced when Dicer was knocked down via siRNA in H1299 cells at a level comparable to what we observed for three known microRNAs (i.e. miR-10b, miR-21 and miR-26; Figure S3A, B), suggesting that the transcript of miR-4423 is processed through the Dicer-dependent microRNA biogenesis pathway. In addition, we examined the data of Hafner *et a*l, (23) who sequenced RNA associated with various components of the AGO complex in human embryonic kidney cells (GSE21918) and found three reads that mapped uniquely to miR-4423-5p and one read that mapped uniquely to miR-4423-3p, indicating that these microRNAs can be incorporated into the AGO complex (Figure S4). The low number of reads mapping to miR-4423 is likely due to the fact that the experiment was not performed in airway epithelial cells where its expression is high. Taken together, our results validate the computational prediction that the expressed sequence is a functional microRNA.

miR-4423 expression is restricted to mucociliary epithelium

The extent of miR-4423 expression in 24 human tissues was assayed using qRT-PCR. Both forms of miR-4423 were expressed in the respiratory tract, with moderate expression observed in the trachea and lung and high expression observed in the nasal and bronchial epithelium (Figure 2A). Using *in situ* hybridization on sections from trachea, mainstem bronchi, and second-generation bronchi from non-smoking donors, we found that the expression of miR-4423 is primarily restricted to the airway epithelium in these tissues (Figure 2B, Figure S5). In addition, we observed expression of both forms of miR-4423 at low levels in the ovary. We hypothesized that this was due to contamination of the ovarian sample with mucociliary epithelium from the fallopian tube. Consistent with this hypothesis, we detected expression of both miR-4423-3p and -5p in an independent sample of total RNA from fallopian tube epithelium (Figure 2A).

The miR-4423 primary transcript (pri-miR-4423) is located approximately 600 bp downstream of *WDR63*, which encodes a subunit of the inner dynein arm complex of motile eukaryotic cilia (24). Due to the close proximity of the two loci, and because the expression of *WDR63* has also been reported in bronchial epithelium (25), we hypothesized that pri-miR-4423 and *WDR63* are coexpressed. To test this hypothesis, qRT-PCR was used to assay the expression of pri-miR-4423, *WDR63* and both mature forms of miR-4423 in the 24 human tissues. We found that the expression of *WDR63* had a strong positive correlation with the expression of pri-miR-4423, suggesting that the two loci are coexpressed and that miR-4423, like *WDR63*, is expressed in the ciliated cells of the airway. Interestingly, the expression of pri-miR-4423 and both mature forms of miR-4423 were also highly correlated across most tissue types with a few exceptions. Most notably, we did not detect expression of miR-4423-3p or -5p in the testis, kidney, placenta and brain, despite detecting high levels of *WDR63* and pri-miR-4423 expression. These data suggest that while pri-miR-4423 is expressed in these tissues, it is not efficiently processed into the mature form (Figure S6A).

Additionally, the expression of miR-4423 together with pri-miR-4423 and *WDR63* were highly induced when normal human bronchial epithelial cells (NHBEs) were differentiated into mucociliary epithelia at an air-liquid interface (ALI), beginning 6 days after the cells were raised to the ALI and continuing through day 13, (Figure S6B). The expression of miR-4423 was also strongly correlated with that of the ciliogenic regulator *FOXJ1* (Figure S7A), and this correlation was stronger than the correlation between miR-4423 expression and that of markers of goblet cells (*MUC5B* and *MUC5AC*), Clara cells (*CC10*) and neuroendocrine cells (*ASCL1*) (Figure S7B). Collectively, these data suggest that miR-4423 is expressed in the ciliated cell population of the airway epithelium.

miR-4423 overexpression results in an increase in the number of ciliated cells at an ALI

In order to determine whether miR-4423 is functionally involved in the development of the airway epithelium, we stably overexpressed and knocked down both forms of miR-4423 in NHBE cells and differentiated them into mucociliary epithelium at an ALI. We found that overexpression of miR-4423 results in a significant increase in the number of cells expressing the ciliated cell markers *FOXJ1* and *β-tubulin* (Figure 3A,B) suggesting that the ectopic expression of miR-4423 is sufficient to promote ciliated cell differentiation in the airway epithelium. However, knockdown of miR-4423 results in only a modest decrease in *FOXJ1* and *β-tubulin*-expressing cells, indicating that either miR-4423 may not be required for ciliated cell differentiation or that the level of inhibition we were able to achieve is not sufficient to induce a miR-4423-deficient phenotype. (Figure S8A, B).

Loss of expression of miR-4423 is associated with lung cancer

As the expression of miR-4423 is increased during the differentiation of mucociliated airway epithelium, we examined its expression in cytologically normal epithelial cells from mainstem bronchial brushings of smokers with (n=5) and without (n=4) lung cancer (see Table S6 for

subject demographics) and found that both the 3p and 5p forms are significantly reduced in smokers with cancer (Figure 4A). Moreover, the expression of miR-4423-3p and -5p, along with the highly correlated expression of pri-miR-4423 and *WDR63,* were also significantly reduced in a large fraction of squamous carcinomas (SCC) and adenocarcinomas (ADC) relative to matched adjacent normal tissues regardless of smoking status (Figure 4B, Figures S9A,B). We extended these observations to the lung cancer RNA-seq dataset from The Cancer Genome Atlas (TCGA) (26), using *WDR63* expression levels as a proxy for the levels of miR-4423 expression, and found that *WDR63* was downregulated in a similar proportion of SCC and ADC (Figure S9C). Interestingly, miR-4423 expression is reduced in squamous metaplasia and is reduced further in SCC compared with normal airway epithelium (Figure S10), indicating that the decrease in miR-4423 expression might be an aspect of a process that occurs early in carcinogenesis. In TCGA data, we did not observe evidence of copy number variation that could account for the loss of *WDR63* expression in lung tumors (Figure S11A, B). We did, however, observe a negative correlation between methylation levels and *WDR63* expression in ADC (but not in SCC) (Figures S12, S13). Collectively, these results suggest that miR-4423 is downregulated in a wide range of lung tumors and in premalignant lesions.

miR-4423 inhibits anchorage-independent growth in lung cancer cell lines

To investigate whether miR-4423 is an inhibitor of a cancer-associated process, miR-4423 was stably expressed in seven SCC cell lines (SW900, H1703, RH2, H2170, Skmes-1, H520 and Calu-1), and in three ADC cell lines (Caul-6, H1299 and A549). Expressing miR-4423 in Calu-6, SW900, H1703 and RH2 cells significantly decreased their ability to form colonies in soft agar (Figure 5A), while it did not affect the anchorage-independent growth capacity of the remaining 6 cell lines. In order to characterize the biological basis of miR-4423 sensitivity with regard to anchorage-independent growth, we profiled baseline gene expression of two cell lines that were sensitive to miR-4423 in soft agar (Calu-6, SW900) and one that was resistant (H2170). We

found that genes that were expressed at higher levels in both miR-4423-sensitive cell lines relative to the miR-4423-resistant line H2170 were significantly enriched in genes important for cell differentiation, cell-to-cell contact/migration, and the cell cycle. In contrast, genes expressed at lower levels were significantly enriched for those involved in apoptosis (Figure S14).

miR-4423 suppresses xenograft tumor growth

Based on these results, we sought to determine whether miR-4423 expression can inhibit lung tumor growth in mouse xenografts. To test this hypothesis, H1703, Calu-6 and H1299 cells stably expressing miR-4423 or a control vector were injected subcutaneously into the backs of immuno deficient mice. We found that while miR-4423 overexpression caused a modest decrease in tumor growth in Calu-6 and H1299 cells (Figure S15), it significantly reduced the size of the tumors formed by the SCC cell line H1703 (Figure 5B). Upon histological examination of the H1703 miR-4423-overexpressing tumors we observed foci of altered morphology. These foci, which were not observed in the control tumors, had a more structured cellular organization (Figure S16). Since E-cadherin plays an important role in cell-cell adhesion and loss of its expression can promote increased tumorigenesis by inducing anchorage-independent growth and colonization of tumor cells (27, 28), we analyzed its pattern of expression in the H1703-derived tissues. We found that the miR-4423-overexpressing tumors contained foci of increased phosphorylated E-cadherin, which were not observed in the control tumors (Figure 5C). This result suggests that miR-4423 overexpression promotes increased formation of cell-cell adhesions and provides a possible mechanism by which it suppresses tumor growth.

Transcriptomic consequences of miR-4423 modulation

To explore the mechanisms by which miR-4423 modulates airway differentiation and lung cancer associated phenotypes, we predicted potential mRNA targets of miR-4423 using the TargetscanS v5.0 (29) and Miranda v3.3a (30) algorithms on 3' UTR sequences obtained from Ensembl and Refseq (Table S7). A total of 809 and 1,578 genes were predicted to be targets of miR-4423-3p and -5p, respectively by both algorithms in both sets of 3' UTR sequences. 181 of these genes

were predicted targets of both forms. While the seed regions of the 3p and 5p forms of miR-4423 are not shared with any other known mammalian microRNAs, we observed a 4-5-nucleotide overlap between the seed region of miR-4423-3p and some members of the miR-449 and miR-34 families (Figure S17A) and a significant number of shared predicted targets (Figure S17B). Interestingly, these microRNA families are conserved across vertebrates, upregulated during mucociliary epithelium differentiation, and are key regulators of multiciliogenesis (4, 31).

To characterize the transcriptomic effects of miR-4423 modulation, we profiled gene expression in H1299 cells transiently overexpressing miR-4423 (n=3) or empty vector controls (n=3) and identified 1,231 genes significantly changed (FDR $q < 0.25$; Figure S18A, Table S8). The set of genes whose expression was reduced upon miR-4423 overexpression was enriched in genes with putative 3p and 5p binding sites ($P = 0.001$; Fisher Exact Test; Table S9). Likewise, genes with putative binding sites for miR-4423 were enriched among genes downregulated with miR-4423 overexpression ($P < 0.001$; KS-Test; Figure S18B).

Genes whose expression was reduced after overexpression of miR-4423 were enriched in chaperone proteins (FDR $q < 0.05$; DAVID), including Hsp70 and Hsp40 family members, which are important in protecting cells from apoptosis and promoting anchorage-independent growth(32). Among the downregulated predicted targets, we found members of signaling pathways critical for anchorage-independent survival and growth of cancer cells, including a catalytic subunit of phosphatidylinositol 3-kinase (*PIK3CA*) (33) and the SH2-containg protein, *SHC1*, which couples activated growth factor receptors to the Ras pathway and promotes cellular transformation (34).

In comparison to previous studies, we found that genes whose expression was altered upon miR-4423 overexpression were concordantly differentially expressed during differentiation of NHBEs at an ALI (FDR $q < 0.05$; GSEA; Figure S19) (35). Consistent with our observation that miR-

4423 expression begins at day 6 of ALI differentiation (Figure S7A), we found that many of the genes altered by miR-4423 overexpression were those whose expression levels began to change specifically between days 4-8 of ALI differentiation (Figure S19). With regards to lung cancer, using a previously published gene expression study from our group (17), we found that genes that decrease upon miR-4423 overexpression are enriched among genes whose expression is increased in airway epithelium from patients with lung cancer (FDR $q < 0.001$; GSEA). We also found in three datasets (36-38) that the genes altered upon miR-4423 overexpression were significantly enriched among genes whose expression changed in the opposite direction in lung ADC and SCC relative to adjacent normal tissue (FDR $q < 0.05$; GSEA; Figure S20A, B). Together, these results suggest that miR-4423 can regulate gene expression changes that occur during both the differentiation of airway epithelium as well as lung carcinogenesis.

Discussion

This study used next-generation sequencing of small RNA from human bronchial epithelium to identify a primate-specific microRNA, miR-4423, which is expressed at high levels in airway epithelium. Although miR-4423 was previously computationally predicted to be a microRNA based on sequencing reads present at extremely low levels in other tissues (21, 22), this study represents the first functional characterization and validation of this microRNA. We have shown that the production of the mature forms of miR-4423 is Dicer-dependent, supporting our assertion that the expressed sequence is indeed a functional microRNA.

The primary transcript of miR-4423 (pri-miR-4423) is situated immediately downstream of the gene *WDR63*, which encodes the human homolog of the dynein intermediate chain IC140 of the ciliated alga *Chlamydomonas reinhardtii* (24) and is therefore believed to be a subunit of the inner dynein arm complex of motile eukaryotic cilia. These two transcripts are coexpressed across a range of human tissues, and both are strongly expressed in multiciliated epithelia (lung, trachea,

and nasal, bronchial and fallopian tube epithelium), and highly induced during airway epithelial cell differentiation in concert with the ciliogenic regulator *FOXJ1*. However, it is intriguing that while both pri-miR-4423 and *WDR63* are highly expressed in the testis, kidney, placenta and brain, mature miR-4423-3p and -5p were not detected in these tissues. Further work is needed to determine the regulation of pri-miR-4423 maturation, and whether there are other cell types in which the expression of pri-miR-4423 is decoupled from pri-miR-4423 processing.

Based on its pattern of expression, we hypothesized that the induction of miR-4423 may be important for the establishment and/or maintenance of mucociliary epithelium. In accordance with this hypothesis, we found that the pattern of gene expression observed upon miR-4423 overexpression significantly overlaps with changes in gene expression observed during the differentiation of NHBEs into mucociliary airway epithelium at an ALI (35). In addition, we found that the overexpression of miR-4423 in NHBEs differentiated into mucociliary epithelium increases the number of cells expressing *FOXJ1* and *β-tubulin*. Knockdown of miR-4423 in ALI cultures led to only a modest decrease in the number of *FOXJ1* and *β-tubulin* expressing cells. An important limitation of this experiment relates to our inability to estimate the efficiency of miR-4423 knockdown, given that our method of inhibiting miR-4423 uses a single-stranded anti-microRNA that inhibits miR-4423 function without altering its expression levels.

Nonetheless, a non-essential role for miR-4423 in airway epithelium differentiation is consistent with its lack of evolutionary conservation, and suggests that this microRNA might act redundantly with other factors. Consistent with this hypothesis, we found that the seed regions and predicted mRNA targets of miR-4423 overlap with those of miR-449/miR-34 family members, which are well-studied regulators of airway epithelial differentiation. However, miR-4423 also has predicted targets that are distinct from those of miR-449/miR-34, and it remains to be determined if miR-4423-dependent modulation of miR-4423-specific targets impacts airway epithelial phenotypes.

We observed that the expression of miR-4423 is decreased in lung tumors, which may simply reflect the absence of ciliated cells in these tissues. However, we also found that the ectopic expression of miR-4423 inhibits anchorage-independent growth *in vitro* and reduces tumor growth *in vivo*. Furthermore, we found that genes that are downregulated by miR-4423 overexpression include members of the PIK3CA and SHC signaling pathways, which are important for anchorage-independent growth. These findings suggest that this microRNA may directly play a tumor-suppressive role. We found that miR-4423 expression is lost in both SCC and ADC. However, given that SCC arises from airway epithelium and the role we found for miR-4423 in airway epithelial differentiation, we hypothesize that loss of miR-4423 expression may be more functionally relevant in SCC than in ADC. Consistent with this hypothesis, both the soft agar and xenograft studies show that miR-4423 has an effect on some of the SCC cell lines tested while it only affected the anchorage-independent growth capacity of the poorly differentiated ADC cell line Calu-6. Additionally, we found that cell lines in which miR-4423 overexpression inhibits anchorage-independent growth show higher expression of genes important for cell differentiation, focal adhesion and cell cycle and decreased expression of genes involved in apoptosis relative to miR-4423-overexpression-insensitive cell lines, suggesting that the loss of miR-4423 might contribute to tumorigenesis in a specific subset of SCC.

Intriguingly, we also found that the expression of miR-4423 is downregulated in the cytologically normal bronchial epithelium of smokers with lung cancer, suggesting that miR-4423 expression might be influenced by a field-cancerization effect. This is further supported by the overrepresentation of predicted miR-4423 targets among genes that are upregulated in the normal airway of smokers with lung cancer, suggesting that the loss of miR-4423 expression may play a role in field cancerization.

These observations, taken together with the potential role of miR-4423 in airway cell differentiation, are consistent with the many examples of processes that contribute to

differentiation and suppress malignancy (39). There are several other examples of tissue-specific microRNAs, such as miR-29 and miR-1/206, whose expression is lost in cancerous tissues and whose ectopic expression promotes redifferentiation and abrogates the malignant phenotype (5, 6). Consistently, histological analysis of miR-4423-overexpressing tumors revealed foci of increased phosphorylated E-cadherin, which were not observed in the control tumors. These results suggest that miR-4423 overexpression can promote cell adhesion and open the possibility that miR-4423 may be capable of restoring aspects of the differentiation program of cancer cells. Additional mechanistic studies will be needed to determine whether and how miR-4423 plays a role in the molecular crosstalk between airway differentiation and lung carcinogenesis.

This work has a number of translational implications for the study of both lung cancer and airway epithelium development. First, miR-4423 is the first example of a microRNA with lung-cancer-associated differential expression in cytologically normal bronchial airway epithelium. This extends our previous findings that changes in mRNA expression in the normal airway are associated with lung cancer and premalignancy (17, 40), to microRNA, and suggests that differences in the expression of miR-4423 and/or other microRNAs might be useful for the early detection of lung cancer in the relatively accessible proximal airway. Second, the effect of ectopic expression of miR-4423 on anchorage-independent growth of lung cancer cell lines and their ability to form tumors in mice suggest the potential of miR-4423 mimetics as lung cancer therapeutics. Third, the species specificity of miR-4423 raises interesting questions about the mechanisms of airway epithelial differentiation and ciliogenesis in simians relative to other mammals, and may have implications for the study of lung injury and other lung diseases in non-simian models. Finally, it is likely that miR-4423 was not discovered before due to its tissue and species-specificity. Therefore, this study demonstrates the potential for unbiased genome-wide profiling to discover novel transcripts with biological and clinical importance.

Material and Methods

A detailed description of the patient populations and methodologies (sample collection, small RNA-sequencing and sequencing data analysis, qRT-PCR, *in vitro* assays, *in situ* hybridization, immunohistochemistry, microarray sample processing and data analysis, air liquid interface cultures and xenograft mouse models) used in this work can be found in the SI Apendix.

Acknowledgements

We thank R. Mallarino, W. Cardoso and Y. Alekseyev (BU Microarray Core) for technical support and advice. This work was funded by R01 CA 124640 (Spira and Lenburg), U01 CA152751 (Spira, Dubinett and Lenburg) as part of the NCI's Early Detection Research Network (EDRN), National Science Foundation Integrative Graduate Education and Research Traineeship (Campbell), P50CA58184 (Belinsky), Merit Review 5I01BX000359 (Dubinett) and R43HL088807-01 (Hayden). All small RNA sequencing and microarray data has been deposited in GEO under the accession number GSE48798.

References

1. Hwang HW, Mendell JT (2006) MicroRNAs in cell proliferation, cell death, and tumorigenesis. *Br J Cancer* 94(6):776-780.
2. Zhao Y, Srivastava D (2007) A developmental view of microRNA function. *Trends Biochem Sci* 32(4):189-197.
3. Lize M, Herr C, Klimke A, Bals R, Dobbelstein M (2010) MicroRNA-449a levels increase by several orders of magnitude during mucociliary differentiation of airway epithelia. *Cell Cycle* 9(22):4579-4583.
4. Marcet B*, et al.* (2011) Control of vertebrate multiciliogenesis by miR-449 through direct repression of the Delta/Notch pathway. *Nat Cell Biol* 13(6):693-699.
5. Fabbri M*, et al.* (2007) MicroRNA-29 family reverts aberrant methylation in lung cancer by targeting DNA methyltransferases 3A and 3B. *Proc Natl Acad Sci U S A* 104(40):15805-15810.
6. Taulli R*, et al.* (2009) The muscle-specific microRNA miR-206 blocks human rhabdomyosarcoma growth in xenotransplanted mice by promoting myogenic differentiation. *J Clin Invest* 119(8):2366-2378.
7. Brase JC, Wuttig D, Kuner R, Sultmann H (2010) Serum microRNAs as non-invasive biomarkers for cancer. *Mol Cancer* 9:306.
8. Kosaka N, Iguchi H, Ochiya T (2010) Circulating microRNA in body fluid: a new potential biomarker for cancer diagnosis and prognosis. *Cancer Sci* 101(10):2087-2092.
9. Shields PG (2002) Molecular epidemiology of smoking and lung cancer. *Oncogene* 21(45):6870-6876.
10. Sridhar S*, et al.* (2008) Smoking-induced gene expression changes in the bronchial airway are reflected in nasal and buccal epithelium. *BMC Genomics* 9:259.

11. Zhang X, *et al.* (2010) Similarities and differences between smoking-related gene expression in nasal and bronchial epithelium. *Physiol Genomics* 41(1):1-8.
12. Beane J, *et al.* (2007) Reversible and permanent effects of tobacco smoke exposure on airway epithelial gene expression. *Genome Biol* 8(9):R201.
13. Chari R, *et al.* (2007) Effect of active smoking on the human bronchial epithelium transcriptome. *BMC Genomics* 8:297.
14. Hackett NR, *et al.* (2003) Variability of antioxidant-related gene expression in the airway epithelium of cigarette smokers. *Am J Respir Cell Mol Biol* 29(3 Pt 1):331-343.
15. Spira A, *et al.* (2004) Effects of cigarette smoke on the human airway epithelial cell transcriptome. *Proc Natl Acad Sci U S A* 101(27):10143-10148.
16. Schembri F, *et al.* (2009) MicroRNAs as modulators of smoking-induced gene expression changes in human airway epithelium. *Proc Natl Acad Sci U S A* 106(7):2319-2324.
17. Spira A, *et al.* (2007) Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. *Nat Med* 13(3):361-366.
18. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10(3):R25.
19. Griffiths-Jones S (2004) The microRNA Registry. *Nucleic Acids Res* 32(Database issue):D109-111.
20. Friedlander MR, *et al.* (2008) Discovering microRNAs from deep sequencing data using miRDeep. *Nat Biotechnol* 26(4):407-415.
21. Stark MS, *et al.* (2010) Characterization of the Melanoma miRNAome by Deep Sequencing. *PLoS One* 5(3):e9685.
22. Witten D, Tibshirani R, Gu SG, Fire A, Lui WO (2010) Ultra-high throughput sequencing-based small RNA discovery and discrete statistical biomarker analysis in a collection of cervical tumours and matched controls. *BMC Biol* 8:58.
23. Hafner M, *et al.* (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* 141(1):129-141.
24. Yang P, Sale WS (1998) The Mr 140,000 intermediate chain of Chlamydomonas flagellar inner arm dynein is a WD-repeat protein implicated in dynein arm anchoring. *Mol Biol Cell* 9(12):3335-3349.
25. Lonergan KM, *et al.* (2006) Identification of novel lung genes in bronchial epithelium by serial analysis of gene expression. *Am J Respir Cell Mol Biol* 35(6):651-661.
26. Deus HF, *et al*. (2010) Exposing the cancer genome atlas as a SPARQL endpoint. *J Biomed Inform* 43(6):998-1008.
27. Asnaghi L, *et al.* (2010) E-cadherin negatively regulates neoplastic growth in non-small cell lung cancer: role of Rho GTPases. *Oncogene* 29(19): 2760-2771
28. Bremnes RM, Veve R, Hirsch FR, Franklin WA (2002) The E-cadherin cell-cell adhesion complex and lung cancer invasion, metastasis, and prognosis. *Lung cancer* 36(2):115-124.
29. Grimson A, *et al.* (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell* 27(1):91-105.
30. Enright AJ, *et al.* (2003) MicroRNA targets in Drosophila. *Genome Biol* 5(1):R1.
31. Wang L, *et al.* (2013) miR-34b regulates multiciliogenesis during organ formation in zebrafish. *Development* 140(13):2755-2764.
32. Khaleque MA, *et al.* (2005). Induction of heat shock proteins by heregulin beta1 leads to protection from apoptosis and anchorage-independent growth. *Oncogene* 24(43):6564-6573.

33. Akca H, Demiray A, Tokgun O, Yokota J (2011) Invasiveness and anchorage independent growth ability augmented by PTEN inactivation through the PI3K/AKT/NFkB pathway in lung cancer cells. *Lung Cancer* 73(3):302-309.
34. Carrano AC, Pagano M (2001) Role of the F-box protein Skp2 in adhesion-dependent cell cycle progression. *J Cell Biol* 153(7):1381-1390.
35. Ross AJ, Dailey LA, Brighton LE, Devlin RB (2007) Transcriptional profiling of mucociliary differentiation in human airway epithelial cells. *Am J Respir Cell Mol Biol* 37(2):169-185.
36. Sanchez-Palencia A*, et al.* (2011) Gene expression profiling reveals novel biomarkers in nonsmall cell lung cancer. *Int J Cancer* 129(2):355-364.
37. Xi L*, et al.* (2008) Whole genome exon arrays identify differential expression of alternatively spliced, cancer-related genes in lung cancer. *Nucleic Acids Res* 36(20):6535-6547.
38. Wachi S, Yoneda K, Wu R (2005) Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. *Bioinformatics* 21(23):4205-4208.
39. Sell S (1993) Cellular origin of cancer: dedifferentiation or stem cell maturation arrest? *Environ Health Perspect* 101 Suppl 5:15-26.
40. Gustafson AM*, et al.* (2010) Airway PI3K pathway activation is an early and reversible event in lung cancer development. *Sci Transl Med* 2(26):26ra25.

Figure Legends:

Figure 1. Expression and evolutionary conservation of miR-4423. The region of human chromosome 1 from nucleotides 85599425-85599600 (human genome build hg19) was visualized using the UCSC Genome Browser. **A)** A coverage plot of the sequencing reads that aligned to miR-4423, scaled to reads per million (RPM). **B)** The boundaries of the primary transcript as predicted by miRDeep, with the 5p and 3p forms highlighted in red. **C)** A multiple species alignment of the human sequence with that of the corresponding genomic regions of 9 nonhuman primates and rat, as obtained from the MULTIZ track in the UCSC Genome Browser. The seed regions of the 5p and 3p forms of miR-4423 are highlighted in red. Matches to the human reference sequence are represented as dots, and gaps relative to the human reference are shown as dashes. A 4-bp insertion in the mouse lemur sequence (relative to the human reference) is indicated with a numeral. GA=great ape, OM=old world monkey, NM=new world monkey, PS=prosimian.

Figure 2. The expression of miR-4423 is primarily restricted to mucociliary epithelium. A) Expression of both forms of miR-4423 across 24 human tissues is detected in the respiratory tract (lung, trachea, nasal and bronchial epithelium), ovary and fallopian tube epithelium. **B)** By *in situ* hybridization, miR-4423 is expressed in the epithelium of the trachea, mainstem bronchus and second-generation bronchus. Arrowheads are pointing to the regions with positive staining.

Figure 3. MiR-4423 overexpression results in an increase in the number of cells expressing ciliated cell markers. A) NHBE cells overexpressing miR-4423 or control were differentiated into mucociliary epithelium at an ALI. NHBE cells overexpressing miR-4423 show substantial *FOXJ1* and *β-tubulin* staining (top) compared to control (bottom). Representative images shown were taken at days 9 (*FOXJ1*) and 11 *(β-tubulin*). Arrows are pointing to regions of positive staining. **B)** Expression of *FOXJ1* was assayed via qRT-PCR in NHBE cells overexpressing miR-4423 or control at different time points of differentiation into mucociliary epithelium at an ALI (Days 1,5,7,9,11,13, 15 and 17). Using a linear model that had *FOXJ1* expression as the response and time point and treatment as predictors we found that the expression of *FOXJ1* is significantly increased in cells overexpressing miR-4423 compared to controls (p=0.025). Expression of *FOXJ1* was normalized using both *eGFP* and *GAPDH*.

Figure 4. MiR-4423 expression is associated with lung cancer. **A**.**)** Expression of miR-4423 was assayed in histologically normal epithelium from the mainstem bronchus of smokers with lung cancer (C)(n=5) and smokers with benign disease of the chest (NC)(n=4). The 3p and 5p forms of miR-4423 are significantly downregulated in the bronchial epithelium of smokers with lung cancer compared to smokers without lung cancer (3p, $P = 0.037$; 5p, $P = 0.045$). **B)** Expression of miR-4423-3p and -5p is significantly downregulated in tumor tissue (T) compared to matched adjacent normal tissue (Adj.N) for: SCC (n=15; 3p, P=0.031; 5p, P=0.01); ADC from current and former smokers (n=10; 3p, P=0.029; 5p, P=0.028) and ADC from non-smokers

(n=10; 3p, P=0.01; 5p. P=0.04). Error bars indicate standard error, and P values were determined using Student's t test in part A and a paired t test in part B.

Figure 5. MiR-4423 inhibits lung cancer anchorage-independent growth *in vitro* and tumor growth *in vivo*. A) Soft agar assays were performed in the indicated cell lines stably transfected with either a vector that overexpresses the miR-4423 precursor or the empty parent vector as a negative control (n=10). Overexpression of miR-4423 in four of the cell lines tested decreases the number of colonies formed in soft agar (Calu-6, P=2.8 x 10^{-7}; SW900, P=2.4 x 10^{-5}; H1703, P=7.4 x 10^{-9}; RH2, P=0.001). Error bars indicate standard error, and P values were determined using Student's t test. **B)** H1703 (SCC) cells stably overexpressing miR-4423 or a control ($1x10^6$) were injected subcutaneously into the backs of NSG mice (7 mice/group). Tumors derived from miR-4423 overexpressing cells were growth suppressed relative to the control-derived tumors as shown in representative photographs of the tumors (left panel), tumor volume over time (P=1.55 $x10^{-11}$) (right top panel) and tumor weight (P=0.01) (right bottom panel). **C)** Phosphorylated E-cadherin staining was performed in miR-4423-overexpressing tumors and controls (H1703). Mir-4423-overexpressing tumors exhibited focal areas of positive membrane phospho-E-cadherin staining consistent with the presence of tight junctions (bottom panel). Phospho-E-cadherin staining was not observed in control tumors (top panel). Arrows are pointing to regions of positive staining.